

Structure in the 3D Galaxy Distribution:

I. Methods and Example Results

M.J. Way^{1,2}, P.R. Gazis and Jeffrey D. Scargle

NASA Ames Research Center, Space Science Division, Moffett Field, CA 94035, USA

Michael.J.Way@nasa.gov, PGazis@sbcglobal.net, Jeffrey.D.Scargle@nasa.gov

1. Abstract

Three methods for detecting and characterizing structure in point data, such as that generated by redshift surveys, are described: classification using self-organizing maps, segmentation using Bayesian blocks, and density estimation using adaptive kernels. The first two methods are new, and allow detection and characterization of structures of arbitrary shape and at a wide range of spatial scales. These methods should elucidate not only clusters, but also the more distributed, wide-ranging filaments and sheets, and further allow the possibility of detecting and characterizing an even broader class of shapes. The methods are demonstrated and compared in application to three data sets: a carefully selected volume-limited sample from the Sloan Digital Sky Survey (SDSS) redshift data, a similarly selected sample from the Millennium Simulation, and a set of points independently drawn from a uniform probability distribution – a so-called Poisson distribution. We demonstrate a few of the many ways in which these methods elucidate large scale structure in the distribution of galaxies in the nearby Universe.

2. Introduction and Historical Background

By the mid-1700s telescopes began to be used to catalog large areas of the night sky. It quickly became clear that the distribution of objects is not homogeneous. Wright (1750) was the first to note that our Sun appears to reside in a disk of stars while Messier (1781) was probably the first to detect a cluster of galaxies. Of the 103 objects in Messier’s catalog 13 are actually part of the Virgo cluster. Of course there was no distinction between galactic and extra-galactic nebulae at this early stage, but an overall inhomogeneity was obvious. In

¹NASA Goddard Institute for Space Studies, 2880 Broadway, New York, NY, 10025, USA

²Department of Astronomy and Space Physics, Uppsala, Sweden

his larger catalog Herschel (1784) discovered the Coma cluster along with voids and other congregations of matter. By 1847 his son John Herschel was able to use his larger catalog of 4,000 nebular objects (Herschel 1847) to quantify the inhomogeneity for the first time using counts-in-cells (15' in Right Ascension by 3° Declination) confirming Messier's discovery of Virgo with the addition of several other clusters and even superclusters of galaxies as we understand them today. Huggins (1864) measurements of nebular spectra would open the door to categorizing these strange objects, but not until 1925 would it be confirmed that the Spiral Nebulae were in fact external to the Milky Way (Hubble 1925) and their distribution on the night sky better understood.

Using the Shapley-Ames, Harvard and Hubble surveys of galaxies in the early 1930s Shapley (1933); Bok (1934); Hubble (1936) and Mowbray (1938) essentially demonstrated that galaxies to at least 18th magnitude are not randomly distributed. Also around this period Hubble (1934) used galaxy counts-in-cells to find for the first time that the distribution of galaxies is log-normal.

By the 1950s the Lick Catalog of galaxy counts (reaching over 1 million and superseding all previous catalogs in scale) could be used to statistically characterize the galaxy distribution. Neyman & Scott (1952, 1959) assumed that “Galaxies occur only in clusters” and built a multi-parameter model to characterize the distribution of galaxies. Then for the first time a number of authors attempted to use the 2-pt correlation function to characterize the galaxy distribution (Limber 1953, 1954; Layzer 1956; Limber 1957; Neyman 1962) using the Lick survey. According to Saslaw (2000), at about the same time “His (Gamow 1954) was probably the first claim that quantitative details of the observed galaxy distribution (Rubin 1954) supported a specific physical theory of cosmogony.”

Characterizing clusters of galaxies from the National Geographic Society – Palomar Observatory Sky Survey (POSS) Abell (1958) used counts in equal-area cells to show that galaxies are more strongly clustered than a Poisson³ distribution. He found the maximum clustering scale to be about 45 Mpc ($H_0=100\text{km/s/Mpc}$), the scale for superclusters. Zwicky (1957) also used the POSS survey but came to the conclusion that clustering stops at the scale of clusters of galaxies and is uniform above that scale. But it was clear from other observations that there *are* superclusters of galaxies (de Vaucouleurs 1953, 1958) present in the local universe.

Using the new Lick Observatory catalog of Shane & Wirtanen (1967) for galaxies brighter

³For reasons described below in §4, we prefer to call such random processes as *uniformly and independently distributed*, more directly indicating their fundamental nature. However, the term *Poisson* is entrenched in much of the literature.

than $m=19$, Totsuji & Kihara (1969) realized for the first time that the two-point correlation function for the spatial distribution of galaxies follows the power-law

$$g(r) = (r_o/r)^s, \quad (1)$$

where r is the distance between galaxies, $r_o = 4.7$ Mpc, and the index s was estimated to be about 1.8. The results were later confirmed by other groups using the same survey (e.g. Groth & Peebles 1977) with very similar results ($s = 1.77$ instead of 1.8, but with the same r_o). Both Martinez & Saar (2001) and Saslaw (2000) do a nice job of reviewing the progress of the use of correlation functions for galaxy distributions. Szapudi & Szalay (1998) is one of the later developments proposing Landy-Szalay (Landy & Szalay 1993) estimators for higher order correlation functions. They claim that it is the most natural estimator (see e.g. Peebles & Hauser 1974).

Turner & Gott (1976) used positions and magnitudes from 1087 galaxies from the Catalog of Galaxies and Clusters of Galaxies (Zwicky et al. 1961) and applied a well defined, objective group identification procedure in contrast to the somewhat subjective criteria used previously (e.g. Holmberg 1937; Reiz 1941; de Vaucouleurs 1975; Sandage & Tammann 1975; Gregory & Thompson 1978). Later these workers applied the same methodology to a small N-body simulation (Turner et al. 1979). In essence they attempted to estimate the surface density of galaxies with volume density enhancements ≥ 10 , as suggested by de Vaucouleurs (1975) at that time. They admitted their catalog would have contamination from foreground and background objects since they did not have redshift information. Nonetheless they assigned 737 galaxies to 103 separate groups and 350 to the field (see Figure 2 in Turner & Gott 1976). The largest group contained 238 members, including Virgo cluster members.

Oort (1983) reviews some of the earliest results on large-scale structure analyses, but also points out the problem with using the increasingly popular correlation function (e.g. Peebles 1980) to characterize all structures in the universe. “The correlation function has proved to be extremely useful in providing such a unified description of the clumpiness. However, it is not suitable for describing the very long filamentary or flat structures that we encounter in superclusters, nor does it describe the large voids between these superclusters.”

The deficiencies of the correlation function led to the use of methods like percolation analysis and Minimal Spanning Trees in the 1980s. For example, Zeldovich et al. (1982); Shandarin (1983); Einasto et al. (1984) were some of the first to attempt to quantify galaxy clustering using percolation analysis. These groups had the belief that it could appropriately quantify the pancake and filamentary structures of the universe in models of structure formation (e.g. Zeldovich 1970). However, Dekel & West (1985) pointed out a number of problems with using percolation analysis and stated that they are in fact not sensitive to the “pancake” structures expected from the calculations of Zeldovich (1970). They recommended a

volume limited sample an order of magnitude denser than the then state-of-the-art Center for Astrophysics survey (Huchra et al. 1983); but even after more dense samples were obtained the validity of the method as a tool for analyzing observational data remained in doubt. On the other hand, it was utilized for comparing N-body simulations with observational data and Poisson (uniform) distributions. More recent percolation work (Pandey & Bharadwaj 2005) has used the SDSS Data Release One (Abazajian et al. 2003) in a 2-D projection to demonstrate that filaments are the dominant pattern in the galaxy distribution.

One now understands the limitations of second-order statistical quantities, such as correlation functions and power spectra, by noting that they discard phase information. As percolation analysis demonstrated the application of more powerful techniques allowing the identification of sheet and filamentary structure in the large scale structure of the universe, at nearly the same time the Minimal Spanning Tree (MST) took hold as a filament-finding algorithm. The MST is a pattern recognition technique borrowed from graph theory which gives an objective measure of the connectedness of a set of points. Barrow et al. (1985) were the first to apply the MST to galaxy clustering using the 2-D catalog of Zwicky et al. (1961), the 3-D catalog of the Center for Astrophysics Redshift Survey (Huchra et al. 1983, hereafter CFA), and the N-body simulations of Gott et al. (1979). These authors demonstrated how markedly different both the observational data and N-body simulations are from a Poisson distribution. Advances in the MST technique have been applied to Large-Scale Structure analysis by a number of other groups in subsequent years (Pearson & Coles 1995; Krzewina & Saslaw 1996; Ueda & Itoh 1997; Doroshkevich et al. 2004; Colberg 2007). The percolation and MST methods are related to Friends-of-Friends (FoF) techniques, which were first applied to the 3-D CFA survey by Press & Davis (1982); Huchra & Geller (1982) and later to simulation data by Croft & Efstathiou (1994) and even larger samples of galaxies to obtain catalogs of groups (Ramella et al. 1997). The FoF technique has even been expanded for use with photometric redshift surveys of galaxies (Botzler et al. 2004). There are additional ways to use the Nth nearest neighbor distances to estimate the underlying density field (e.g. Gomez et al. 1998; Dressler 1980). Another approach is to use all N nearest neighbors (Ivezić et al. 2005) within a Bayesian probability framework.

It should surprise no one that wavelets, used to characterize structure in large galaxy catalogs, were applied in other 2-D (e.g. Slezak et al. 1990) cases, and in the 3-D case (e.g. Slezak et al. 1993). What is surprising is that they have not been utilized more extensively in the largest modern redshift surveys of galaxies (e.g. Martinez et al. 2005). Paredes et al. (1995) have done a nice job of comparing the relative merits of MST, FoFs and wavelets as cluster finding algorithms, although there have been significant developments since.

By the late 1980s and early 1990s there was interest in attempting to measure the topol-

ogy of Large Scale Structure from observational data and various models (Gott et al. 1986; Hamilton et al. 1986; Gott et al. 1987; Park & Gott 1991; Beaky et al. 1992). This was done using the genus statistic which is related to the fourth Minkowski functional (Stoyan et al. 1985). These kinds of measures should give an idea of the topological connectedness of a systems of points after they have been smoothed by some kind of filter. In the end this method allowed one to distinguish among different galaxy distributions by obtaining the genus, using isodensity surfaces at different density levels. These clearly require some kind of smoothing, but the choice of levels at which to apply smoothing is not obvious. This is important because over-smoothing tends to create a positive genus, while under-smoothing creates a negative one. Nonetheless these problems have not stopped groups from applying these techniques to the largest available redshift surveys of galaxies available at the moment, such as QDOT, CfA2, PSCz, 2dFGRS, and the SDSS (Moore et al. 1992; Vogeley et al. 1994; Canavezes, et al. 1998; James et al. 2007; Gott et al. 2009). Sheth et al. (2003) used Minkowski Functionals combined with percolation analysis to compare the supercluster-void network in three cosmological models and that of the present epoch. Some of the latest studies (Gott et al. 2009; Choi et al. 2010) seem to confirm a sponge like topology, and is consistent with the Gaussian random phase initial conditions expected from inflation. Recent work (Aragón-Calvo et al. 2010; Zhang et al. 2010) has attempted to calculate Minkowski Functionals using Delaunay Tessellation to calculate the isodensity surfaces to try and get around the smoothing problem mentioned above.

Voronoi tessellation was applied for the first time to study the structure of the universe with the pioneering works of Matsuda & Shima (1984) and Icke & van de Weygaert (1987). This was extended to 3-D distributions by Yoshioka & Ikeuchi (1989) and van de Weygaert (1994). In the meantime Voronoi tessellation-based methods have been used to study the clustering of galaxies by many for differing purposes (e.g. Coles 1990; Ikeuchi & Turner 1991; Kim et al. 1999; Ramella et al. 1999, 2001; Pizarro et al. 2006; Aragón-Calvo et al. 2007). For example, Ebeling & Wiedenmann (1993), used a high-density selection in the distribution of Voronoi volumes, coupled with the adjacency information, to develop a method for source detection in 2D point maps. This approach has been adapted into analysis toolkits for Chandra X-ray source identification; see *e.g.* Diehl & Statler (2006) for details. Melnyk, Elyiv & Vavilova (2006) applied a similar threshold method to study the distribution of 7,000 local supercluster galaxies. See Elyiv, Melnyk & Vavilova (2009) for discussion of an extension of Voronoi tessellation to more complex neighbor relationships. See Cappellari (2009) regarding various applications. Two of our methods utilize this procedure, and details are found below in §4 and §5.2.

The pace of development of innovative methods for charactering large scale structure has not much diminished in recent years. Two recent methods first generate a contin-

uous density field from the 3-D point distribution and then identify structures via similar means. Aragón-Calvo et al. (2007) use the “Delaunay Tessellation Field Estimator” (Schaap & van de Weygaert 2000; Schaap 2007) and then rescale using isotropic Gaussian filters to create the continuous field, while Bond et al. (2009) use a fixed-width Gaussian kernel to estimate the density field. They both then compute the matrix of second spatial derivatives to yield the so-called Hessian matrix. The eigenvalues and eigenfunctions of this continuous matrix are evaluated at the locations of the galaxies yielding clouds of points in what Bond et al. (2009) call λ -space. Bond et al. (2009) demonstrate the relationship between the shapes of these clouds and the morphology of the corresponding structures – clusters, sheets, and filaments in particular. Aragón-Calvo et al. (2007) use what they term the “Multiscale Morphology Filter” which “looks to synthesize global structures by identifying local structures on a variety of scales and assembling them into a single scale independent map”. Aragón-Calvo et al. (2007); Jones et al. (2010) convincingly demonstrate the abilities of their technique via toy models, complex N-body simulations and the SDSS. The Bond et al. (2009) technique is unlike adaptive smoothing (e.g. Stein 1997), because Bond et al. (2009) smooth separately on a series of length scales, with the goal of characterizing the spatial structures more accurately. Choi et al. (2010) use a Hessian approach to compare the length of filaments found at a redshift of ~ 0.8 to 33 lower-redshift subsamples from the SDSS to find that the length scales have not changed very much over this range of redshifts. van de Weygaert & Schaap (2009) review in excellent detail the use of density estimation in “The Cosmic Web” via the “Delaunay Tessellation Field Estimator”. After submission two other papers (Sousbie 2010; Sousbie, Pichon & Kawahara 2010) using DTFE as a density estimator were submitted which characterize the cosmic web and filamentary structure using a method from computational topology called Morse theory.

Recently Hahn et al. (2007a,b) have developed a classification scheme designed to distinguish between dark matter halos in four structures; clusters, filaments, sheets and voids, in N-body simulations of the universe. The scheme relies upon the dynamical differences of the four different structures quantified by an application of the Zeldovich (1970) approximation to the evolved density field which allows one to determine their asymptotic dynamics. There is one free parameter that acts as a smoothing parameter for the density field. Nonetheless they claim to be capable of quantifying the redshift evolution of dark matter halo properties of mass and environment. This is comparable to work by a number of authors in recent years (e.g. Lemson & Kauffmann 1999; Sheth & Tormen 2004; Croton et al. 2007).

While characterizing the clustering of galaxies was the initial focus of many researchers void characterization in 3-D simulations and surveys has also been of interest. Recently Colberg et al. (2008) assembled 13 different void-finding algorithms and for the first time tested them all on a single data set – the Millennium Simulation (Springel et al. 2005).

They claim that the results agree very well with each other. Since then two other interesting approaches with zero or few free parameters have appeared. Platen et al. (2007) have utilized the watershed transform to develop what they term the “watershed void finder” to find voids in 3-D distributions in a “relatively” parameter free way (also see Sousbie, Colombi & Pichon (2009)). Neyrinck et al. (2005); Neyrinck (2008) have used Voronoi tessellation to develop a relatively parameter free “halo-finding” algorithm called VOBOZ (VORonoi BOund Zones) and another to find voids and subvoids called ZOBOV (ZOnes Bordering On Voidness) “without any free parameters or assumptions about shape”.

Regardless of method, clusters and voids were clearly visible in the first large area redshift survey: The Center for Astrophysics Redshift Survey (Huchra et al. 1983) and explicitly described in Davis et al. (1982). Davis et al. (1982) also discuss the discrepancies between their observational data and N-body simulations⁴ at the time: “We also present redshift-space maps generated from N-body simulations, which very roughly match the density and amplitude of the galaxy clustering, but fail to match the frothy nature of the actual distribution”.

Giovanelli & Haynes (1991) has an excellent summary of the largest redshift surveys up to 1991, by which time there were approximately 30,000 galaxies with measured redshifts. Surveys up to 1990 were mainly done with single slit spectrographs in the optical or 21-cm H I line surveys of spirals and gas-rich dwarfs, both measuring one galaxy at a time. Since that time the number of measured galaxy redshifts has increased by orders of magnitude because of advances in large format CCD technology in combination with multi-fiber and multi-object spectrographs. One of the first of these new surveys was the Las Campanas Redshift Survey (LCRS Shectman et al. 1996) which collected over 23,000 redshifts in 6 years. As one can surmise from the above historical survey of methods, it was expected that a large variety of techniques would be applied in rapid fashion by a large number of groups. For example, Doroshkevich et al. (1996) applied a “core sampling technique” (Buryak et al. 1994) to find the characteristic scales for large scale structure in the LCRS. A few years later Doroshkevich et al. (2001) combined inertia tensor and minimal spanning tree analysis to three-dimensional data to confirm their earlier LCRS results and determine cluster dimensions.

The next large redshift survey completed was the Two Degree Field Galaxy Redshift Survey (Colless et al. 2001), which collected approximately 250,000 galaxy redshifts. The state of the art at present is the Sloan Digital Sky Survey (York et al. 2000) with over 1 million measured redshifts thus far, with more on the way.

⁴20,000 points, 150Mpc on a side via Efsthathiou & Eastwood (1981)

The availability of these new large-area low-redshift surveys has greatly enhanced prospects for an objective quantitative description of so-called large scale structure (LSS) as delineated by optical and other observations of galaxies. In addition to the intrinsic importance of assessing large scale structure itself, links between structure and galaxy morphology or color have provided much of the inspiration for a explosion of interest in large-scale observational surveys.

In fact there are several near-future large-area surveys of the sky which will allow one to test the predictions of general relativity for the growth of structures in the universe and its consistency with the history of cosmic expansion (e.g. Stril et al. 2010; Rapetti et al. 2009). A sampling of these surveys include the Large Synoptic Survey Telescope (LSST) (Ivezic et al. 2008), PanStarrs (Kaiser et al. 2002), and BigBOSS (Schlegel et al. 2009).

One of the oldest uses of large scale structure analysis is in the area of the *environmental effects* on galaxy formation and evolution. Starting from the time of Hubble (1936) astronomers have found that the properties of galaxies are dependent upon conditions in their surroundings. Since then a large and varied research effort has explored the dependence of galaxy color, morphology, and star formation history on local density, using ever larger samples of galaxies (e.g. Oemler 1974; Butcher & Oemler 1978; Dressler 1980; Postman & Geller 1984; Santiago & Strauss 1992; Zehavi et al. 2002; Hogg et al. 2003; Kauffmann et al. 2004; Croton et al. 2005; Blanton et al. 2006; Blanton & Berlind 2007; Zehavi et al. 2010).

Part of the present work differs from the tessellation procedures referenced above by combining Voronoi cells into contiguous sets, called *blocks*, using a statistically principled method called *Bayesian blocks* (Scargle 1998, 2002; Scargle et al. 2008). The blocks are collected into contiguous sets to form structures meant to model the shapes of clusters and other large scale entities. Since no constraints – such as spherical symmetry, convexity, or even simple-connectivity – are imposed on the derived structures, our results are useful for detecting and characterizing complex structures such as filaments, sheets, and irregular clusters, not just classical galaxy clusters. This approach is consonant with the notions of the *Cosmic Web* and *Voronoi Foam* (van de Weygaert 2003; van de Weygaert & Aragón-Calvo 2009). Although we leave analysis of the detection efficiency for such complex structures to the next paper in this series, the flexibility of the *Bayesian blocks* representation of the density field allows such structural features to be detected and characterized

Our approach to density estimation is outlined in Section 3, the data sets used are described in Section 4, density and structure estimation methods in Section 5, results in Section 6, and conclusions in Section 7.

3. Basic Approach: Density Estimation plus Structure Analysis

The approach here is the commonly adopted one of treating galaxies as mass points,⁵ using positional and redshift data from surveys to determine locations of these points in three-dimensional space. As described below the subsequent structure analysis flows from the coordinates of the points themselves, and by determining the properties of a postulated underlying continuous field.

Several factors impose limits on this approach. First, note that the data are inherently four, not three, dimensional: distant galaxies are placed by the data where they were a look-back time prior to now, not where they are now. Interpretation of any data analysis results must account for this lack of co-temporality.

Next, there is an inevitable positional uncertainty due to random observational errors in the basic data and systematic effects arising in the transformation from redshift to spatial coordinates. For example, see the discussion of redshift distortion in §18.2 of Saslaw (2000).

And finally note that there are fundamental limitations on the information that can be extracted from coordinates of a set of points. One can carry out statistical analysis directly on the discrete data points, for example by studying multiple-point correlation function estimators, the distribution of nearest neighbor distances, the related minimal spanning trees, and the like. Another, more or less complementary approach, is to postulate the existence of an underlying continuum field, and regard the points as samples related in some way to the field. However, the meaning of such a continuum is problematic in general, especially at small spatial scales – *e.g.* less than that characterizing galaxy nearest neighbor separations.

One such continuum scheme is to regard the field as an estimate of the density of points (say in units of galaxies per cubic parsec), smoothed on scales at least as large as the typical distance between points, and very much larger than the sizes of the galaxies, which are after all treated as points of zero size. Excellent overviews of the mathematical aspects of multivariate densities and their estimation from point data are to be found in Silverman (1986); Scott (1992). Discussions of this concept in relation to the large-scale structure of the Universe are found in Martinez & Saar (2001); Saslaw (2000); Dekel & Ostriker (1999).

A different, but related, scheme interprets the field as a probability distribution, and treats the galaxies as points drawn from it in the usual statistical sense. More specifically, this process can best be viewed as a doubly-stochastic process, sometime called a Cox pro-

⁵Throughout, the terms *galaxy* and *point* will be used more or less interchangeably

cess. The spatial dependence of the galaxy formation is described by process 1, reflecting the evolution of the initial density fluctuations into a formation rate parameter in a probability distribution locally defined in space-time. Process 2 represents the random sampling from the rate determined by process 1. That is to say, the actual appearance of a galaxy in the data is a second random process, independent of the first, reflecting the appearance of a galaxy at a given point in space-time. Indeed, one could separate the galaxy formation and observational detection aspects into two separate, independent processes, if such a triply stochastic representation should prove useful. A mathematical introduction to the basics of such random processes can be found in Papoulis (1965), and excellent overviews of the mathematics of the corresponding theory and estimation methods are Snyder (1991); Daley & Vere-Jones (2002); Andersen et al. (1992); Kutoyants (1998); Preparata & Shamos (1985); de Berg et al. (1997).

Both of the above approaches have to deal with difficult problems related to the fact that the points are not independently distributed with respect to both processes 1 and 2, due to the physics of the underlying formation, evolution, and clustering processes and observational effects (such as the “fiber collision” problem described below). These and other issues are well described in a large literature (e.g. Martinez & Saar 2001).

All of the algorithms used in this paper have some relation to density estimation from points. But some go farther. For example, spatial Voronoi or Delaunay tessellations extract information about relations between galaxies – in terms of quantities such as local galaxy density gradients, nearest neighbor distances (where, importantly, the number of nearest neighbors is not fixed, but rather determined by the data themselves), the distributions of these distances, and information about connectivity within the galactic network that forms the skeleton of the Cosmic Web.

4. The Data

We have applied our three techniques (based on adaptive kernel smoothing, self-organizing maps, and Bayesian blocks), to three individual datasets (one observed, one simulated, and one a simulated purely random distribution).

Dataset 1 is a volume limited sample drawn from the SDSS DR7 (Abazajian et al. 2009, DR7) Main Galaxy Sample (MGS) Catalog (Strauss et al. 2002) which contains a redshift for each galaxy. The dataset was drawn from the DR7 in the same manner that Cowan & Ivezić (2008, hereafter CI08) generated their sample from the SDSS data release 5 (Adelman-McCarthy et al. 2007). We chose to use the DR7 sample because the sample

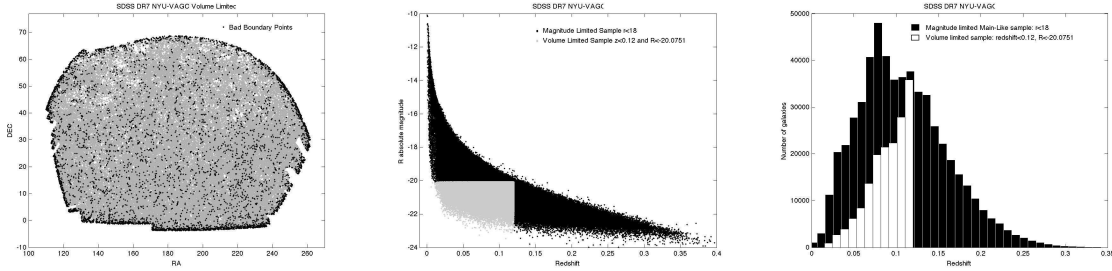


Fig. 1.— Views of the SDSS DR7 data. Left: Positions of galaxies in the Volume Limited (VL) selected SDSS DR7 catalog showing the boundary points that are removed. Middle: The full SDSS DR7 and the volume limited sub-sample selection. Right: Redshift histograms of the full SDSS DR7 and Volume Limited samples.

is larger and essentially geographically contiguous in the north galactic cap region. Rather than use the standard SDSS *casjobs* interface to obtain the actual data⁶ the New York University Value Added Galaxy Catalog (NYU-VAGC) (Blanton et al. 2005) was utilized. The NYU-VAGC includes the k-corrections for all galaxies from the MGS spectroscopic survey. This makes generating the volume limited sample rather trivial. Figure 1 shows the selection of the volume limited subset of the NYU-VAGC sample, after a selection of apparent magnitude in $r < 18$ which mimics the MGS properly. Figure 1 also shows the respective redshift distributions of the Magnitude Limited and Volume Limited Samples.

The MGS sample is obtained from the SDSS via the `primtarget` flag: `primtarget=TARGET_GALAXY (p.primtarget & 0x00000040 > 0)`. The photometric quality is constrained via the three flags `!BRIGHT` and `!BLENDED` and `!SATURATED`: $((\text{flags} \& 0x8) = 0)$ and $((\text{flags} \& 0x2) = 0)$ and $((\text{flags} \& 0x40000) = 0)$, respectively. All redshifts are required to have an SDSS defined redshift confidence better than 0.95 ($z\text{Conf} > 0.95$) and there should be no redshift estimation warning errors ($z\text{Warning} = 0$). Our sample contains 561,421 galaxies at this stage. An example of what the query would look like in *casjobs* is given in Appendix A. The query shown does not include the absolute magnitudes or k-corrections, as these were obtained from the NYU-VAGC catalog.

The SDSS also has a fiber collision issue which will play a role for density estimation. In essence, fibers cannot be placed closer than $55''$ to each other. However, overlap of repeated plates in some areas means that in fact redshifts have been measured for both galaxies in many pairs separated by less than $55''$. To eliminate bias and ensure a homogeneous sample, we removed a randomly chosen member of each such pair.

⁶<http://casjobs.sdss.org>

Our volume limited data set was drawn from the 561,421 galaxies in the NYU-VAGC DR7 data set above. The largest contiguous region in the South Galactic Cap was chosen and then a redshift/color cut of $z < 0.12$ and $M_R < -20.0751$ was applied yielding 146,112 galaxies (see Figure 1). These samples were then processed as follows:

1. Generate angular (2D) separation information: Find each galaxy’s 6 nearest neighbors on the sky. We verified that this process guarantees identification of all neighbors within $55''$. Deleting randomly chosen members in these close pairs eliminated 6,314 galaxies from the sample.
2. From redshifts and sky coordinates generate 3D Cartesian coordinates, in redshift units, for each remaining galaxy.
3. Generate 3D nearest neighbor information by calculating distances to the 12 nearest neighbors. This number was chosen for convenience, to avoid statistical issues that might be associated with a smaller number of neighbors. This neighbor information was used only in the self-organizing map approach.
4. Generate the Voronoi tessellation of the remaining set of galaxies. This yields the cell vertices associated with each galaxy, from which one finds the identities of the variable number of near neighbors in the Voronoi-Delaunay sense.
5. Calculate from the tessellation information a set of derived parameters, including the cell volume V and radius $R_{Voronoi}$, defined as $(\frac{3V}{4\pi})^{1/3}$; the distance d_{CM} between each galaxy and the center of its cell; and an ‘elongation’ measure equal to the ratio between the maximum and minimum dimension of the cell (See Appendix B).
6. Normalize the nearest neighbor distances and the Voronoi radius ($R_{Voronoi}$) by the radius $d_{uniform} = 3.2 \times 10^{-3}$ associated with a uniform density distribution. This information was used in both the self-organizing map (SOM) and Bayesian block (BB) approaches. Scale also the offset distance d_{CM} by $R_{Voronoi}$.
7. Flag questionable samples: Apply a set of tests to eliminate Voronoi cells that appear to be distorted by boundary effects. These tests are described in detail in a discussion of the ‘Boundary Problem’ in section 5.2.2. 5807 points are removed which is about 4% of the initial volume limited sample of 146,112.

After the removal of the boundary points and those within $55''$ of each other we are left with 133,991 points.

Combining these derived data (nearest neighbor distances and characteristics of Voronoi cells) with attributes taken directly from the survey data (positions, photometry data, etc.) yielded a unified set of attributes for each galaxy as described in Appendix B below.

Dataset 2 is a volume limited sample drawn from the Millennium Simulation (Springel et al. 2005, hereafter MS). We follow the same recipe for creating our sample as is done by CI08 to make it comparable to the SDSS sample. After a redshift and magnitude cut to mimic the SDSS Main Galaxy Sample ($r < 18$ and $0.005 < z < 0.25$) there are 509,877 galaxies. Another redshift and absolute magnitude cut is made to mimic the SDSS volume limited sample described above ($R < -20.0751$ and $z < 0.120$). This leaves 171,388 galaxies in our simulated volume limited sample. See Figure 2 for a representation of these samples.

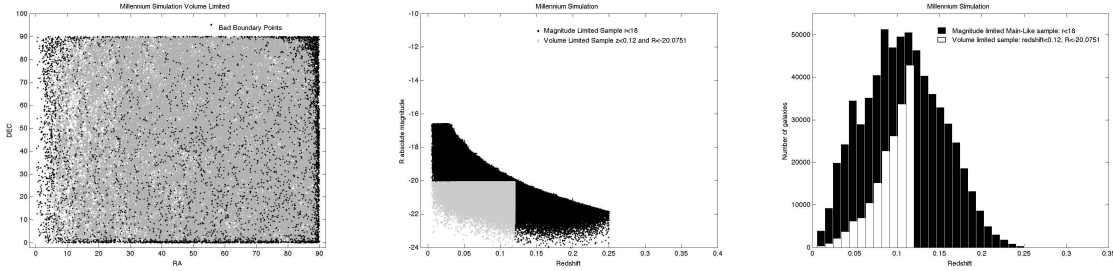


Fig. 2.— The data from the full Millennium Simulation displayed as in Figure 1: Left: Positions of galaxies in the Volume Limited (VL) selected Millennium Simulation catalog showing the boundary points that are removed. Middle: The full Millennium Simulation and the volume limited sub-sample selection. Right: Redshift histograms of the full Millennium Simulation and Volume Limited samples.

Dataset 3 is a set of randomly distributed points that mimics the SDSS DR7 Volume Limited sample above. We took a cube of space enclosing a volume equivalent to the SDSS DR7 Volume Limited sample. We then filled this cube with points drawn independently from a spatially uniform probability distribution. It is common to call this a Poisson distribution, because the number of such independent and uniformly distributed points in a predefined volume of size V obeys the Poisson distribution, $N(n) = (\lambda V)^n e^{-\lambda V} / n!$, where λ is the event rate per unit volume. It can be confusing to use the same term for this auxiliary distribution as for the overall spatial distribution. We therefore prefer to call the random process based on its essential nature: independent, or for the case where the rate parameter λ is constant, independent and uniform. (Indeed, the “Poisson” nature of this distribution is merely an incidental consequence of these properties.) The number of points was chosen such that, after removing pairs just as with the SDSS fiber collision criterion (none closer than $55''$), there remained a number of galaxies (144,700) close to that in the SDSS DR7 Volume limited

sample. Note that this sample differs from the others in two separate ways: the uniformity of the distribution and its simple, geometrical boundary. For the most part the former is the more important consideration.

5. Structure Estimation Methods

As described in the Sections 2 and 3, analysis of large scale structure is not a simple matter, especially if one wishes to invoke an underlying continuum. Here we describe the various methods we have used, each of which explores a different aspect of the distribution of galaxies on various scales.

5.1. Kernel Density Estimation

Kernel Density Estimation is probably the most widely used non-parametric density estimator in use today. For this reason several groups have used 3D kernel density estimation in recent years to study the large scale structure of the Universe from redshift surveys (e.g. Connolly et al. 2000; Balogh et al. 2004), and we include such an analysis in order to compare the results of our two newer methods to this well known approach.

The underlying idea of 3D kernel density estimation (KDE) is simple: construct a 3D profile (or *kernel*) centered at each data point, and sum the contributions of these kernels for all of the data points. The kernels and their sums are evaluated at a grid of 3D points, typically arranged in a uniform rectangular grid. What needs to be specified are: the shape of the kernel (Gaussian and Epanechnikov kernels are commonly used) and its width⁷ (this can be fixed or adaptive to the underlying distribution) and amplitude, plus the locations of the grid elements.

Since our other two methods are effectively adaptive (although the adaptivity is implemented differently), we use an adaptive-bandwidth Gaussian kernel to calculate the density. To describe it as simply and transparently as possible we first explain the 1D univariate case and then 3D. In 1D one first starts by estimating the density with a fixed bandwidth (h) where the Gaussian kernel (K) is given by Equation 2. Equation 3 is then the density estimate (p) for the 1-D fixed bandwidth case where the points are given by x_i . To estimate the variable or adaptive 1D KDE one allows the bandwidth to vary from point to point. Let $d_{i,j}$ represent the distance from point x_i to the k th nearest point in the set making up the

⁷Sometimes called *bandwidth*, although strictly speaking this term refers to the frequency domain.

other $n - 1$ data points. Equation 4 represents the 1D variable KDE where one sees that the window width of the kernel at point x_i is proportional to $d_{i,j}$ such that regions with sparser data points will have flatter kernels. Hence the new adaptive bandwidth could be represented as $h_i = h \times d_{i,j}$. This estimation method is based on the approach laid out by Silverman (1986).

In the 3D case one has to find an initial estimate of the density for each point, normally by using the fixed bandwidth 3D KDE shown in Equation 5. One then must build a local bandwidth term λ_i at each point. These should have unit (geometric) mean and be multiplied by the global bandwidth h . In this case h is the overall smoothing and λ_i adjusts the bandwidth at each point to “adapt” to the density of the data. The 3D adaptive density estimate is given by Equation 6.

However, multi-dimensional multi-bandwidth KDE on large data sets can be computationally expensive. In order to deal with a large number of points (e.g. 100,000) in a reasonable time Gray & Moore (2003a,b) have devised an efficient “Dual Tree” algorithm. The algorithm also gives an error within a user specified tolerance at any evaluated point. Rather than code the algorithm ourselves we utilized a package of MatLab⁸ routines based on the Kernel Density Estimation Toolbox of Ihler⁹ which has implemented the dual tree algorithm of Gray & Moore (2003a,b). We made some small modifications to allow the code to run on 64-bit platforms so that one could evaluate the largest of our data sets.

$$K = e^{-\frac{(x-x_i)^2}{2h^2}} \quad (2)$$

$$p(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (3)$$

$$p(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{hd_{i,j}} K\left(\frac{x-x_i}{hd_{i,j}}\right) \quad (4)$$

$$p(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_h} K\left(\frac{x-x_i}{h}\right) \quad (5)$$

⁸© The Mathworks, Inc.; <http://www.mathworks.com>

⁹<http://www.ics.uci.edu/~ihler/code>

$$p(x) = \frac{1}{n} \sum_{i=1}^n \frac{1}{V_h \lambda_i} K\left(\frac{x - x_i}{h \lambda_i}\right) \quad (6)$$

The Kernel Density Estimation (KDE) method gives an almost continuous distribution of densities. In order to make easier comparisons between this and the two other methods to be discussed below we have translated the continuous distribution of densities into discrete classes. This was done by collecting the base-10 logarithms of the densities into a small number of bins. For the SDSS DR7, Millennium Simulation, and uniform random data sets this led to 11, 13, and 10 KDE logarithmic density classes, respectively, chosen to approximately match the SOM-based class structure.

5.2. Tessellation

Tessellation is a natural partitioning scheme for analysis of the distribution of points in a space of any dimension. We have found it exceptionally useful for this study of the spatial distribution of galaxies. Accordingly, two of our structure analysis procedures (Bayesian blocks and self-organizing maps) use as building blocks the elements of the Voronoi tessellation of 3D space defined by the galaxy positions, as described in the following subsection.

5.2.1. Voronoi Tessellation

Tessellation divides the data space into sub-volumes, here called *cells*. The first four of the following are properties of tessellation in general, while the last two are specific to Voronoi tessellation in three dimensions (Okabi et al. 2000):

1. N data points generate N cells.
2. The cells and data points are in a one-to-one correspondence.
3. The union of all N cells is the whole data space.
4. The intersection of any pair of cells is empty (no cell overlap).
5. A cell comprises that part of the data space closer to its data point than to any other.
6. The cell boundaries are flat 2D polygons.

7. Computation of the tessellation yields a data structure containing the following information:

- (a) An estimate of the local point density: V^{-1} , where V is the cell volume.
- (b) The 3D vector from cell centroid to data point estimates the local density gradient, in both magnitude and direction.
- (c) Information on nearest neighbors is encoded in the vertices of the bounding polygons. One can define two cells to be *adjacent* in three ways, depending on whether they share at least one vertex, edge, or face; in this order, each definition is included in the next.

In regions of high density, a small volume is apportioned among many points, so the cells are small. In low density regions, where points are few and far between, the opposite is true: the cells are large. This is the key inverse relationship between density and cell size (*cf.* item 5 in the list in §4), supplemented by the gradient information

Each cell is that part of the data space dominated by the corresponding data point (item 5); in Voronoi tessellation, this means in the sense of being closer to it than to any other data point. Items 3 and 4 together mean that the tessellation is a *partition* of the data space. The subsidiary information in item 7 exemplifies the way in which both point and local information are conveniently represented in the tessellation construct. Our Bayesian block and self-organizing map schemes make direct use of this information in different ways, as described in later sections. In the former case density and geometrical information alone is used to gather cells into connected sets, called *blocks*, to represent the underlying density structure. In the latter case incorporation of other subsidiary information allows the SOM representation to describe more general characteristics of the large-scale structure.

In both cases, the adjacency information encoded in cell faces, edges and vertices is rather like a list of nearest neighbors – where the number of neighbors is not pre-set, and in fact is part of the information extracted from the raw data. Further, the density gradient information mentioned above can be utilized for analysis and for visualization purposes. A handy density visualization scheme depicts each cell as a frustum with the Voronoi cell as the base with straight vertical sides, and capped by a copy of the Voronoi cell at a height $\rho_i = n_i/V_i$ where the number of points (often 1) is divided by the cell volume. This fast and convenient density representation involves no loss of information by binning or smoothing, but therefore has a discontinuous and ragged appearance. Display issues limit this device to data spaces of dimension 1 or 2 (and therefore it is not used here); nevertheless this construct is useful for computing subsidiary quantities such as widths of structures, local mean density

gradient, *etc.* In short Voronoi tessellation yields a convenient data representation that enables many useful local, intermediate, or global quantities to be computed.

There are many excellent, fast algorithms for tessellating spaces of any dimension. We used the Matlab routine Qhull (Barber et al. 1996) which is computationally efficient and returns adjacency and other auxiliary information in a convenient form. Without any further computations, the Voronoi cells express considerable statistical information about the point distribution. For example, Figure 3 shows the distribution functions of the local densities computed as the reciprocal of the volumes of the Voronoi cells for the three cases: the SDSS DR7 data, the Millennium Simulation data, and the uniform data. These distributions characterize the dynamic range of the cell sizes. As expected, the cells in the uniform case have a relatively narrow distribution centered around the mean cell size, while in the other cases a broader range reflects the presence of structure on a wider range of scales. The degree to which the distribution for the case of the MS data is similar to that for the DR7 data confirms the correctness of this aspect of the simulations. While the log densities are approximately normally distributed, the density distributions themselves have long tails that render the (log) of the mean value a misleading central measure.

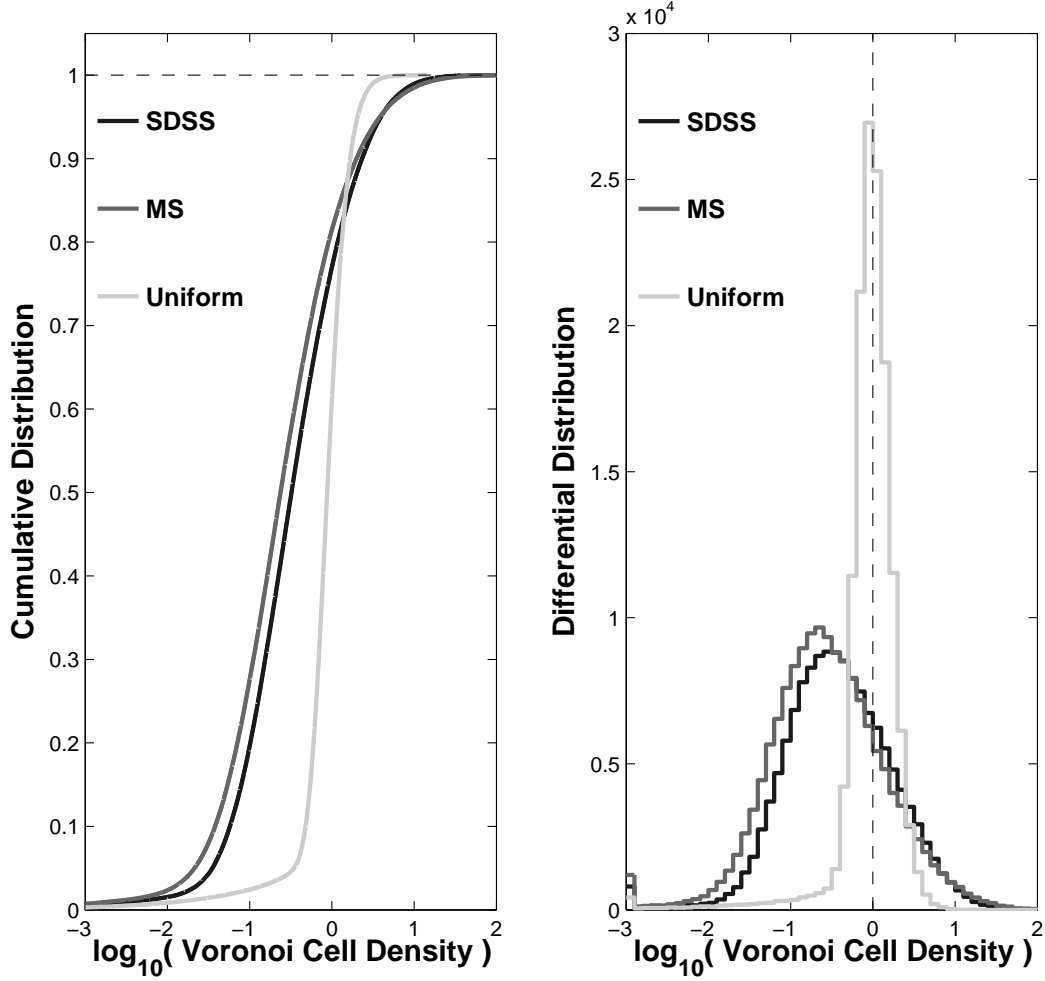


Fig. 3.— Distribution functions of the logarithm of local densities, computed as the reciprocals of the volumes of each galaxy’s Voronoi cell. In both panels: dark line = SDSS DR7, medium line = Millennium Simulation, light line = spatially uniform random distribution. Left: unbinned cumulative distributions. Right: differential distributions. All distances (r) used to calculate the volumes are in redshift units (z): $r(z) = 3 \times 10^3 h^{-1} z$ Mpc. The units of volume for the random uniform case are chosen so that the mean is unity (indicated by the vertical line at $\log(\text{cell density})=0$).

Figure 4 compares the distributions of the number of neighbors of each cell. A neighbor of a cell is defined to be any cell sharing one or more Voronoi vertices with the given cell. In this case the distributions of the actual DR7 data and the MS simulation data are nearly indistinguishable, whereas that of the random data is distinctively different.

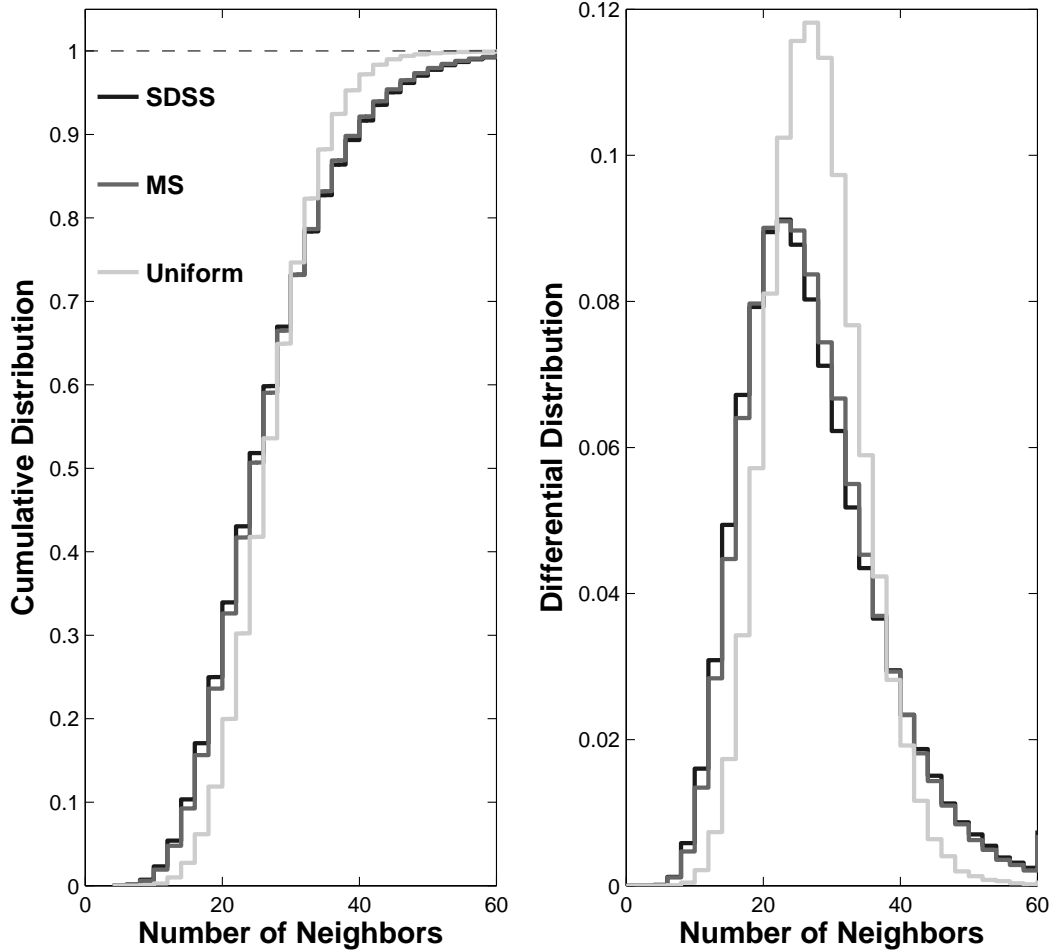


Fig. 4.— Normalized distribution functions of the number of Voronoi neighbors of individual galaxies. In both panels: dark line = SDSS DR7, medium line = Millennium Simulation, light line = spatially uniform random distribution. Left: unbinned cumulative distributions, normalized to unit total fraction. Right: differential distributions.

Figure 5 depicts the distribution functions of the logarithm of the average distance to the Voronoi neighbors of each galaxy. As expected, the actual and simulated galaxy data shows much more dispersion than does that for the randomized case.

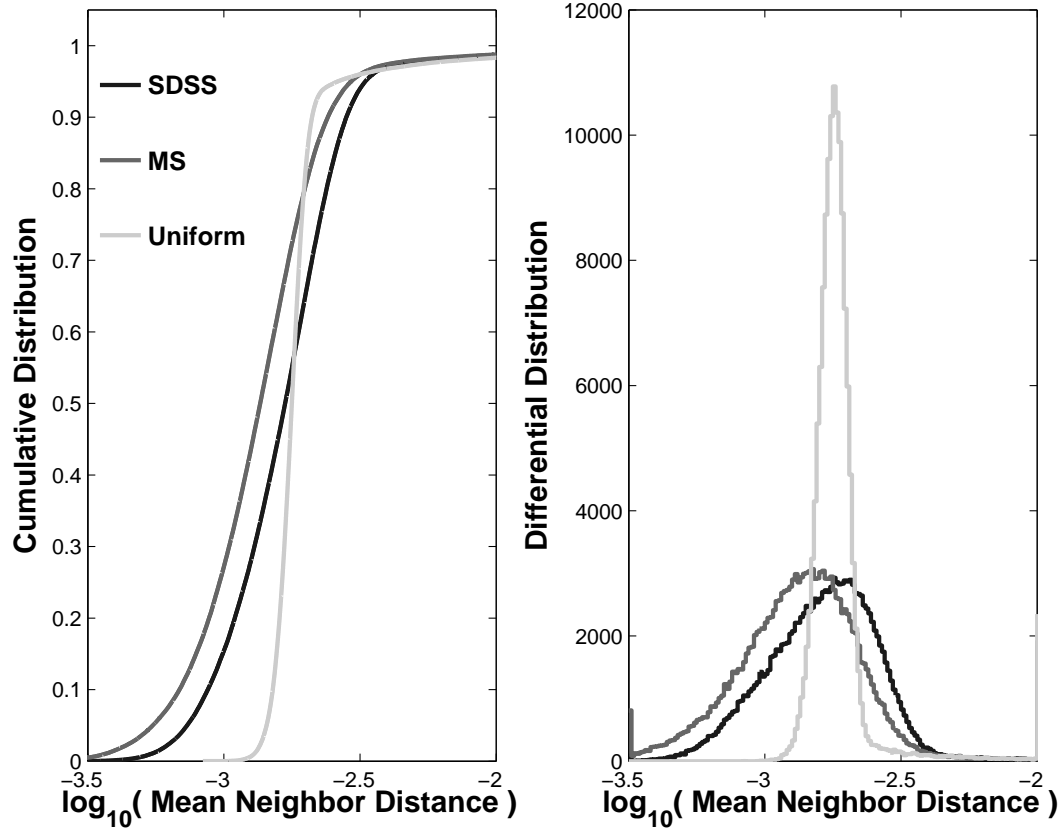


Fig. 5.— Distribution functions of (log) mean distances to Voronoi neighbors. In both panels: dark line = SDSS DR7, medium line = Millennium Simulation, light line = spatially uniform random distribution. Left: unbinned cumulative distributions, normalized to unit total fraction. Right: differential distributions.

5.2.2. *The Voronoi Cell Boundary Problem*

For points lying sufficiently deep within the main population Voronoi tessellation is a stable and well-understood procedure that gives meaningful results. For galaxies near an edge of the sample space the situation becomes problematic. Some cell vertices for these points characteristically lie unrealistically far beyond the sampled region. Such outsized cells are an artifact due entirely to the sampling and not to the actual galaxy distribution. For this reason and other difficulties, such as vertices formally assigned to lie at infinity, the reliability, or even the meaning, of the tessellation as a density estimation tool breaks down near the edges of the volume populated by the data points. This is the Voronoi Tessellation ‘Boundary Problem’.

It is possible to attempt to fix the problem, either by modifying the Voronoi tessellation procedure itself or by modifications to the data set. One possibility would be to construct replacement data cells, truncated to finite volumes, as surrogates for the offending cells. However, unless the edges of the sample space are well defined and smooth, procedures of this sort tend to be arbitrary, and can introduce problems of their own. For a data set bounded by complex boundaries with irregularly-shaped indentations and projections there is no simple way to distinguish every cell that suffers from the Boundary Problem from those that do not without eliminating a larger than necessary number of points. Note that after the submission of our paper a similar study to ours was also submitted (Sousbie, Pichon & Kawahara 2010). They deal with the boundary problem in the SDSS in a relatively simple manner by defining boundary points as those that “belong to a pixel with at least one completely empty neighbor”. While we agree that this method is simple and effective, we believe it removes too many non-boundary points and given the already small size of our volume limited sample we did not feel this would be appropriate.

Regardless, it is possible to devise a set of ad hoc criteria that will identify all of the worst case situations without excluding a prohibitive number of ‘good’ samples. These criteria were obtained by studying the distributions of various parameters of the Voronoi cells, in order to set corresponding thresholds.

We evaluated a wide range of different parameters by using the complete data set and subsets of the data that filled simple convex shapes. This was used to help determine which parameters tended to assume extreme values for samples at a boundary without excluding an unacceptable number ($N < 1-200$) of the samples well inside the data volume (what we call the ‘interior region’). The boundary points were identified by the extent of their Voronoi cells with respect to the edge. The parameters most sensitive to the position of a sample with respect to a boundary were $R_{Voronoi}$, d_{CM} , and the normalized distance from the center of a Voronoi cell to its furthest apex, R_{Max} . We used these three parameters in conjunction

to obtain the best performance. We evaluated our criteria for a range of different thresholds to verify that the results were comparatively insensitive to the values of these thresholds. The final values used are listed in Table 1.

The choice of the ‘interior region’ mentioned above is described as follows:

1. One desires a region deep enough inside the full sample region such that one is certain that no sample in this interior region will suffer from the ‘boundary problem’. To ensure this one has to be certain that even samples with extremely large Voronoi volumes have volumes that lie inside the full sample region.
2. An ‘interior region’ is chosen with a boundary that lies $10 \times d_{uniform}$ inside the boundary of the full sample region. Recall that $d_{uniform} = 3.2 \times 10^{-3}$ in units of redshift.
3. To extend outside the full sample region a point in this ‘interior region’ would have to have at least one dimension of its Voronoi volume greater than $10 \times d_{uniform}$ in length. If the volume was shaped as a very thin slice (which is unlikely) it could reach to the boundary, but our own tests showed that this did not take place in our data sets. Regardless, this means that the volume would be roughly $(10 \times d_{uniform})^3$ and our tests show that the number of samples with volumes that size or larger in our interior region is extremely small: $N < 1-200$ as mentioned above.
4. One can conclude that an interior region with a boundary $10 \times d_{uniform}$ inside the boundary of the full data set cannot contain a significant number of points that suffer from the boundary problem.

The number of affected boundary data points was small (our selection criteria flagged 5807 of 146112 points or $\sim 4\%$ of the population), so we simply mark them to exclude them from any further analysis.

5.3. 3D Bayesian Blocks using Voronoi tessellation

This section describes the modeling procedure we used for the 3D galaxy distribution using the Bayesian blocks algorithm. In a nutshell, we partition the data space with a set of surfaces enclosing 3D solids. A constant density is assigned to each solid which is equal to the number of galaxies within it divided by its volume. This partitioning is implemented via an optimization procedure designed to express spatial density variations that are real, and at the same time suppress statistical fluctuations that are not real. The former is regarded as the true signal and the latter as noise (especially that due to the presence of small numbers

of points). Of course these two goals cannot be achieved perfectly. The corresponding signal-to-noise tradeoff is mediated by the model fitness function (detailed below in §5.3.2). As in 2D there are an infinite number of ways to partition a given volume. However, allowing only partitions whose elements are collections of the polyhedra defined through the Voronoi tessellation of the data points, as described in §5.2, yields a completely tractable, finite, combinatorial optimization problem.

In summary, the goal of finding the optimal piece-wise constant model is achieved with the Bayesian block algorithm. Optimality is in the sense of maximizing a measure of goodness-of-fit of models of this kind. The basic elements, *i.e.* the Voronoi cells, are determined using standard computational geometry algorithms. In the next subsections we describe how the cells are collected together into density levels, and how the cells within a level are collected together to form connected blocks. The assembly of blocks into meaningful structures (such as clusters, sheets, filaments, or other structures) will be described only briefly, as details will appear in a separate paper.

5.3.1. Levels

The segmentation process described above begins by collecting the galaxies into levels – *i.e.* sets forming a hierarchy ordered by density (galaxies per unit volume). The goal is to find the best piecewise constant model described above (§5.3). This optimization is implemented with an algorithm Jackson et al. (2010) that maximizes goodness-of-fit for piecewise constant models. This procedure for optimal segmentation of a data space of any dimension is an extension of a one-dimensional algorithm Jackson et al. (2005) that in turn is an exact, dynamic programming based version of the approximate algorithm in Scargle (1998).

In general a set of 3D, or even 2D, data cells cannot be ordered in a way that allows implementation of the basic idea behind the 1D algorithm.

Extension to higher dimension Scargle (2002); Jackson et al. (2010) is achieved by discarding the condition that the elements of the partition of the data space be connected sets of cells. That is to say, the levels are generalized to be arbitrary subsets of the cells in the tessellated data space. Since relaxing this constraint slightly changes the fundamental problem and results in a larger search space, it would seem to be counterproductive. It turns out that the resulting simplicity of the problem outweighs the enlargement of the search space. Without the contiguity constraint the actual locations of the cells are irrelevant to the model. Accordingly all orderings of the cells are equivalent. It is convenient to sort

them in a 1D array ordered by cell volume. Now if the fitness function satisfies a simple convexity condition each level in the optimal 3D partition contains all the cells in an interval in the ordered 1D cell array, and only those cells. It is this “intermediate density” order property that allows the 1D algorithm to find the optimal partition of the original 3D data. The convexity condition referred to is that the fitness function is convex as a function of the number of galaxies in the block and also of block volume, and has nothing to do with convexity of the block or level structures. See Jackson et al. (2010) for details.

One problem results from this approach: the partition elements, here called *levels* in analogy with the contour levels in topographical maps, are typically fragmented into a number of disconnected parts – much as cartographic contours for the same level can be disconnected. The next section describes our treatment of this issue: in a nutshell identify the parts of each level that are indeed connected, and use these as the building blocks for large scale structure.

5.3.2. *Blocks*

The innovation of our approach, compared to previous Voronoi tessellation methods is that neighboring cells are collected together into levels and blocks (structures within which the galaxy density is modeled as constant) in a statistically principled way. A block is a set of cells constrained to be connected, but not restricted to have any particular shape properties such as convexity or simple connectivity. Various abstract definitions of *connectedness* are used in topology, but with finite spaces the basic ideas are simple: a connected set consists of one piece, not two or more disconnected pieces; a simply connected set additionally has no holes. More formally, in a *connected* set any pair of cells in the block can be joined by a path consisting of an ordered list in which each successive pair of cells are touching. This is sometimes called *path-connected*. In a *simply connected* set, the same is true, but in addition there are no cases where a pair of cells is joined by two or more paths that cannot be smoothly distorted into each other.

Since the blocks represent coherent structures of sensibly constant galaxy density, it is natural to associate them with astrophysically meaningful structures. Without implying any assumption about structural evolution or gravitational binding, we assume that our blocks do correspond to coherent structures in the galaxy distribution.

As presaged in the previous section one ramification needs to be discussed: A given optimal level may well consist of a set of *disconnected fragments* – sets of one or more cells spread throughout the data space and not touching each other. To the extent that a

partition’s levels are not connected, it does not solve the constrained optimization problem originally posed.

If it turns out that each level has only one such component (*i.e.* is simply connected), then *de facto* we have solved the original problem. The levels would then be regarded as the connected blocks that we originally sought. But if not, then what? If some levels consist of two or more fragments detached from each other, it is easy enough to identify these fragments and re-label them as separate blocks. One can consider the resulting partition an approximate solution (to the constrained problem) or as an exact solution of a related problem of equal or greater astrophysical interest (the unconstrained problem). The analog presented by topographical maps, with contour lines indicating loci of constant altitude, may serve to clarify. Suppose that the altitude values are assigned based on some statistical measure, and not fixed at even multiples or the like. Then there would be two choices, namely to constrain or not constrain distinct closed contours to be assigned the same value. That is to say, use a global vs. a local statistical measure to determine contour values. The results presented below incorporate this *post facto* re-labeling of block fragments as blocks.

To fully define the optimization problem we need to specify a quantity to be maximized, such as a goodness-of-fit measure for the piece-wise constant block model. That is, we maximize a measure of how well the data in a given block are modeled as points randomly and independently distributed (with a single constant probability density) uniformly across the block. A number of such fitness functions were described in Scargle (1998), but here we use a maximum-likelihood based fitness function described in Scargle et al. (2008), namely the logarithm of the maximum likelihood for a model, of a block of volume V containing N points in which the event rate is constant.

Before exhibiting this fitness function, a few comments are in order regarding the nature of the random process we are postulating for each block. Our idealized mathematical picture is that the spatial locations of events (galaxies) within the block have two properties:

Independence: the occurrence of an event at any location does not affect the occurrence of any other event at any location.

Uniform distribution: The probability of an event occurring in any given block does not depend on where in the block the interval lies.

Note that these conditions are stronger than the usual, weaker assumption that the events are uncorrelated: independence implies uncorrelated, but not vice versa. However, neither of these conditions is rigorously true. In addition to observational issues, such as the fiber collision effect, the physical process of galaxy formation prohibits the formation of two galax-

ies at the same location. We are relying on this kind of correlation being important only at small scales compared to those under study here. On the other hand, the distribution of galaxies is of course not actually constant over significant spatial regions. In this sense, we are simply forming the best piece-wise constant (or step-function) approximation to a distribution that is presumably continuously variable.

Hence, as in Scargle (1998) for time series data, we are led to model the points in a block as identically and independently distributed with a single probability that is constant across the block. As mentioned above this process is often called a *constant rate Poisson process*, because under it the number of points in a fixed volume obeys the *Poisson distribution*:

$$P(N) = \frac{(\lambda V)^N e^{-\lambda V}}{N!} \quad (7)$$

giving the probability P that N points fall in volume V , when the event rate is λ events per unit volume. The usual derivation of this formula as the limit of repeated Bernoulli trials (see e.g. Papoulis 1965) has led to a common misunderstanding that it is fundamentally an approximation, but the above equation is exact – absent correlations of the sort discussed above.

Maximizing the expression in equation (7) leads to the following maximum likelihood fitness function for the block model of the full data interval:

$$L_{max} = \prod_{k=1}^K \left(\frac{N_k}{V_k} \right)^{N_k} e^{-N_k} \quad (8)$$

where N_k is the number of points in block k , V_k is the volume of block k , and the product is over all blocks in the model, covering the whole observation region (Scargle et al. 2008). The corresponding logarithmic fitness for a block, as implemented in our algorithm, is simply

$$\log L_k = N_k \log \frac{N_k}{V_k} \quad (9)$$

for each block, and

$$\log L = \sum_{k=1}^K N_k \log \frac{N_k}{V_k} \quad (10)$$

for the total model comprising K blocks. In the last two expressions a term proportional to N_k is dropped because, when summed over k , it contributes an unimportant constant to the fitness of the full model. Note that these likelihood expressions depend on only the *sufficient statistics* N and V , and not on the actual distribution of the points within the interval. This fact – somewhat counterintuitive, as this quantity is meant to measure the goodness-of-fit of the assumed uniform distribution – follows because under our model only the total number of events, and not their locations, matters.

In the semi-Bayesian formalism of this model, the fitness function must be augmented with a term that expresses prior information about the value for K , the number of blocks. Optimization using equation (10) without such a supplement tends to yield a large number of blocks, as many as $K \approx N$. Specification of a *prior probability distribution* $P(K)$ is the Bayesian approach to this model complexity problem. A convenient choice for favoring a small number of blocks is the geometric prior:

$$P(K) \sim \gamma^{-K} , \quad (11)$$

where γ is some constant. If the log of this prior is added to the fitness of each block, the appropriate prior is assigned to the model for the full interval. While it is not a smoothing parameter, its value regulates the number of blocks, in effect influencing the apparent smoothness of the representation. In most cases the details of the block representation do not change much for a broad range of values of $\log(\gamma)$, and derived quantities (such as the sizes of structures) tend to be even less sensitive to the adopted value of $\log(\gamma)$. The main departure from a rigorous Bayesian analysis is the fact that K , while weighted according to the prior distribution described above, it is not explicitly marginalized, but instead is optimized in a dynamic programming algorithm.

Figure 6 shows the density levels for the DR7 data, organized by level and block. There are three densities that can be assigned to a given galaxy (here denoted cell n)

1. the cell density: N_{cell}/V_{cell}
2. the block density: N_{block}/V_{block}
3. the level density: N_{level}/V_{level}

where N_{cell} is the number of galaxies in a cell n , here always unity, N_{block} is the number of galaxies in the block containing cell n , and N_{level} is the number of galaxies in the level containing cell n . The cell, block and level volumes are defined in an obvious and similar way. In the figure, the ordinate is the block density of the individual blocks, and the horizontal lines indicate the level density assigned to all of the blocks in that level. Note the lack of overlap of block densities from one level to the next, a result of the algorithm.

5.3.3. Galaxy Structures: Sets of Blocks

Fruitful analysis of the galaxy density distribution can be carried out directly from the blocks themselves, without regard to aggregation into structures. Indeed, the same is true

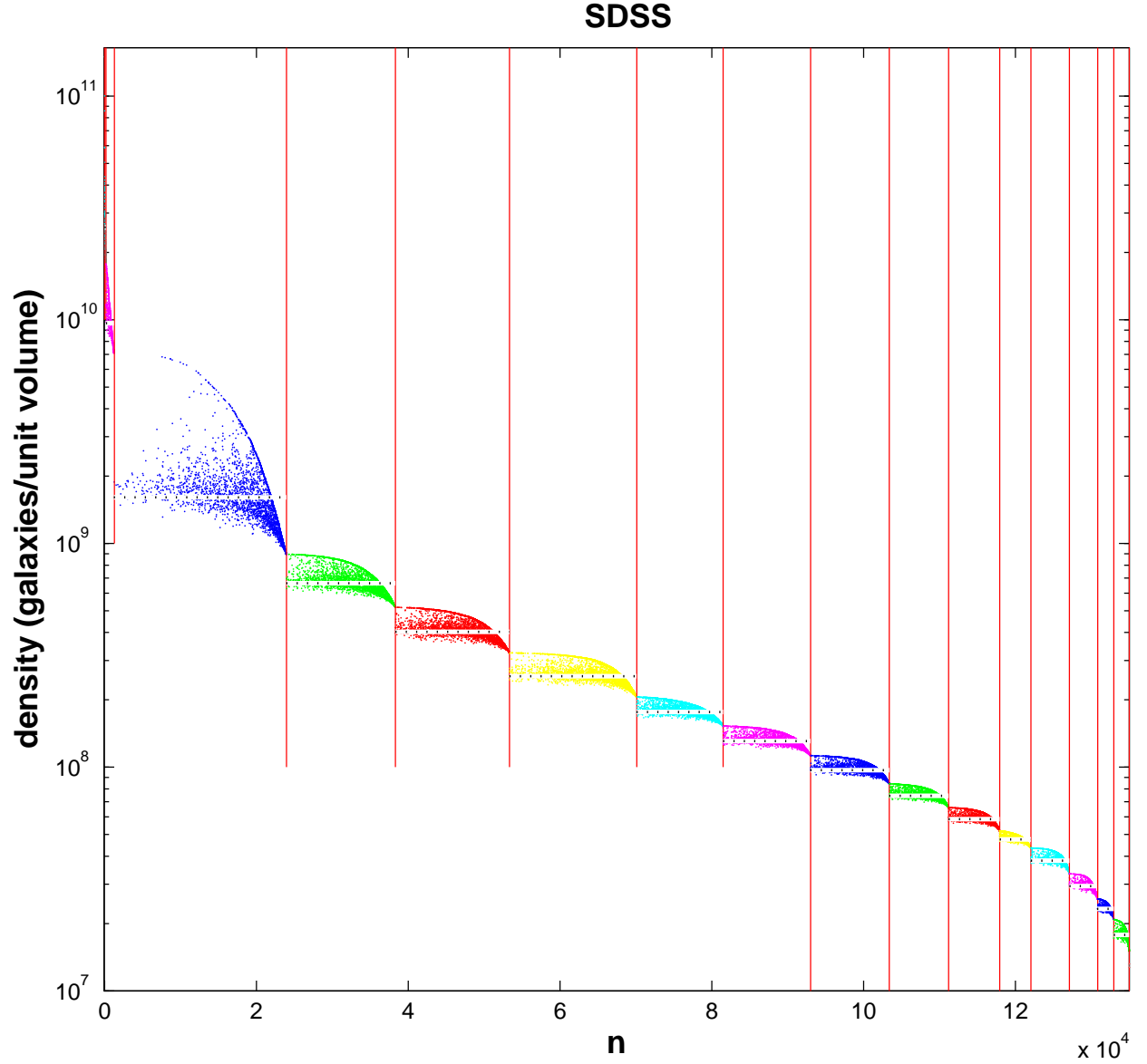


Fig. 6.— Pictorial representation of the density values associated with the different levels (shown in different colors) and blocks within the levels. The base-10 logarithm of the density estimate – number of galaxies per unit volume in redshift units cubed – is plotted against an arbitrary index ordered by level. (The order within the levels is not meaningful. In particular, the curved structure of the envelopes of the points is merely due to the order in which the algorithm identifies blocks within the level.) The horizontal dashed lines indicate the mean galaxy densities in the levels. The distribution is truncated at the bottom-right end for display purposes.

even at the level of Voronoi cells. However, for various applications and for comparison with other work oriented toward cataloging clusters, voids, etc., it is useful to take the aggregation process one step farther and collect neighboring blocks with different densities together to form structures – not just clusters in the classical sense, but also filaments, sheets, and other coherent structures.

Of the many possible algorithmic approaches to this step, we adopt a straightforward approach. First identify local *density maxima*: blocks with a higher density than any block adjacent to it. In 3D there are three ways of defining adjacency: blocks can be deemed adjacent if they share Voronoi cell (1) vertices, (2) edges, or (3) faces. Almost no difference in the deduced structure results from using these progressively restrictive definitions, and throughout we use definition (1).

Next, consider these maxima as seeds, growing into larger structures by attachment of adjacent blocks in the next lower level in the density hierarchy. This procedure is repeated until terminated by some stopping condition. Three examples are: (a) stop at a fixed level in the density hierarchy, either locally (for each structure) or globally; (b) stop when the structure contains blocks for a fixed number of levels; and (c) stop when all blocks belong to one cluster or another. In void analysis, one would adopt a similar strategy beginning at the lower end of the density hierarchy. This approach has some resemblance to that of Platen et al. (2007). In the preliminary large-scale structure analysis reported here we adopt version (b), taking the structures to consist of the block defining the local maxima plus blocks from the two next lower density levels.

5.4. Self-organizing maps

Self-organizing maps (SOMs) (Kohonen 1984; Ritter et al. 1992) are widely used for unsupervised classification. They map points in the input N -dimensional data space \mathcal{R}^N into an array of cells or principal elements (PEs) in a classification space \mathcal{A} of reduced dimensionality (usually one or two dimensions). The algorithm is designed to make the output of the SOM reproduce, as much as possible, the topological structure of the input distribution. In particular it attempts to map adjacent clusters in the input space into adjacent PEs (or more commonly, adjacent blocks of contiguous PEs) in the output space. A variety of measures have been proposed to evaluate the degree to which topology is preserved by a particular mapping (Villmann et al. 1997; Bauer & Villmann 1997; Hsu & Halgamuge 2003).

Used alone, SOMs serve as a means to visualize complicated relationships between

groups of points. For classification purposes, they must be combined with some partitioning scheme that can identify regions in the output map that correspond to different clusters in the input data. We used a modified version of the same *Bayesian Blocks* algorithm described for direct cluster analysis in §5.3 (Scargle 1998; Scargle et al. 2008; Jackson et al. 2010) to partition SOMs. This algorithm partitions the SOM output space into contiguous segments (*blocks*) in a way that optimizes a fitness function which measures how constant the values of the attributes are within each segment.

Let the array of attributes (two in our case) in principal element i of the SOM output map be denoted x_i , and the corresponding variance measure by σ_i^2 ; then the relevant average attribute for block k is

$$X_k = \frac{\sum_i \frac{x_i}{\sigma_i^2}}{\sum_i \frac{1}{\sigma_i^2}}, \quad (12)$$

where the summations are over the N_k PE’s in block k . The fitness function for block k takes the form (Gazis & Scargle 2008)

$$C_k = (N_k - 1)(\ln(R) + \ln \sqrt{\pi}) - (\ln(\prod_i \sigma_i) + \ln(\sum_i \frac{1}{\sigma_i^2})) - (\sum_i \frac{x_i^2}{\sigma_i^2} - X_k), \quad (13)$$

where again the sums are over the PEs in the block. The cost for the entire partition is

$$C = \sum_{k=1}^K C_k. \quad (14)$$

In the SOM case the space to be partitioned is the map itself and the blocks will consist of clusters of contiguous PEs. Note that this is subtly different from the conventional Bayesian Blocks approach, in which partitioning is performed in the original data space.

SOMs were generated using the Neuralware package, discussed at length by Merényi (1998). This software can use a variety of neighborhood schemes and implements the ‘conscience’ algorithm proposed by DeSieno (1988) to prevent any particular PE from representing too much of the input data. Classifications were performed using a 7×7 array of PEs. Neighborhoods were rectangular, and decreased in size from 5×5 to 1×1 during training. Multiple classifications were performed using different values for the range and standard deviation parameters in Equation (13) to evaluate the sensitivity of the algorithm to these parameters. These partitionings were also compared with the best possible partitioning and the results of a conventional threshold-based scheme.

One advantage of SOM-based classification is that it can be performed on any set of parameters. In principle kernel density and Bayesian Blocks methods could be modified to include other parameters, but for a SOM this extension is natural – essentially automatic.

Care must be taken to chose parameters that are physically meaningful. Initially we tried using the $N + 1$ nearest neighbor distances as a proxy for N -point correlation functions, but the results were too sensitive to statistical fluctuations that occur when N is small. Our final classifications were performed using two parameters: a scaled Voronoi radius, $R_{Voronoi}/d_{uniform}$, and an offset distance, $d_{CM}/R_{Voronoi}$, where

$$R_{Voronoi} = (3V_{Voronoi}/4\pi)^{1/3}, \quad (15)$$

$V_{Voronoi}$ is the volume of the Voronoi cell of that galaxy, and $d_{uniform}$ is the average spacing between points in an independent uniform distribution. These parameters are good proxies for the mean and gradient of the local density, respectively. *Bagging* (short for bootstrap aggregating) was performed to improve accuracy and stability, avoid over-fitting, reduce variance, and provide estimates of the uncertainty of the SOM classifications. This standard machine learning procedure involves running the complete analysis algorithm on data sets comprising subsamples from the actual data in the bootstrap fashion (randomly sample with replacement). We averaged the results of 10 such randomly selected subsets of the full data set.

The SOM-based scheme partitioned the SDSS and Millennium Simulation (MS) data sets into six classes. The SOM based scheme partitioned the uniformly random data set into eight classes (see Table 4), but given the non-physical nature of these classes they were not easily defined and will not be discussed further. Based on inspection of the SDSS and MS spatial distributions we identified the six SOM classes as: *Cluster*, *Cluster Gradient*, *Strong Gradient*, *Field Gradient*, *Halo* and *Field*. We indicate these fundamental classes, the number and identity of which are determined by the SOM, in italics. Roman type is used for the names of the somewhat less fundamental BB and KDE classes, derived by clumping their fine-grained densities in order to approximately match the populations of these SOM classes, as detailed in §6.1. This classification could be further refined into eight sub-classes: *Dense Cluster*, *Cluster*, *Dense Cluster Gradient*, *Cluster Gradient*, *Strong Gradient*, *Field Gradient*, *Halo* and *Field*. It should be noted that this later partitioning was determined entirely by the distribution of two attributes ($R_{Voronoi}/d_{uniform}$ and $d_{CM}/R_{Voronoi}$) used by the SOM, and did not involve any a priori choice of thresholds to identify particular categories. The characteristics typical of galaxies in the classes were determined by a *post facto* inspection of the results and summarized in Table 2.

Among the different bagged data sets the boundaries of the six main classes were almost identical; while the subclasses were less consistent their general structure was preserved. Attempts to probe deeper into the hierarchy did not produce stable results, which suggests that any structure that might exist at deeper levels is ambiguous and/or poorly-determined.

The six classes identified by the SOM algorithm can be characterized as follows. The

Table 1: Boundary Tests

Attribute	SDSS		Millennium Simulation		Uniform	
	Threshold	Number ¹	Threshold	Number ¹	Threshold	Number ¹
$R_{Voronoi}$	0.0040	4147	0.0040	4904	0.0040	3556
d_{CM}	0.0023	4515	0.0023	3475	0.0023	5001
R_{Max}	0.0067	5566	0.0067	6022	0.0067	6636
Union ²	-	5807	-	6178	-	6649
Fraction ³	-	0.0415	-	0.0398	-	0.0480

¹Number that failed this test.

²Number that failed one or more of the 3 tests.

³Fraction of samples that failed one or more of the 3 tests.

Table 2: Classes Identified by the SOM Algorithm, Ordered by Mean Density. Note that these class ID numbers only apply to the SDSS and MS datasets. See §5.4 for details on these classes.

ID	Class	Subclass	Characteristics
1	<i>Cluster</i>	<i>Dense Cluster</i>	very high density, low gradient
1	<i>Cluster</i>	<i>Cluster</i>	high density, low gradient
2	<i>Cluster Gradient</i>	<i>Dense Cluster Gradient</i>	very high density, moderate gradient
2	<i>Cluster Gradient</i>	<i>Cluster Gradient</i>	high density, moderate gradient
3	<i>Strong Gradient</i>	<i>Strong Gradient</i>	very high gradient
4	<i>Field Gradient</i>	<i>Field Gradient</i>	moderate-high gradient
5	<i>Halo</i>	<i>Halo</i>	moderate density, low-moderate gradient
6	<i>Field</i>	<i>Field</i>	low density, low-moderate gradient

Cluster class involved regions of high density and low gradient associated with centers of clusters. The *Halo* and *Field* classes involved regions of moderate and low density respectively, with low gradient. Samples were distributed uniformly in space, though galaxies in the *Halo* class may have had some tendency to be associated with the outer portions of clusters. The *Cluster Gradient* class involved regions of high density and moderate gradient associated with filaments and the outer portions of clusters. The *Strong Gradient* and *Field Gradient* classes involved regions of extremely high gradient and moderate gradient, generally of high density, associated with the portions of filaments midway between clusters. The *Field Gradient* class involved regions of low density and moderate gradient respectively, with moderate to low density, and were associated with filaments. This is illustrated by Figure 7.

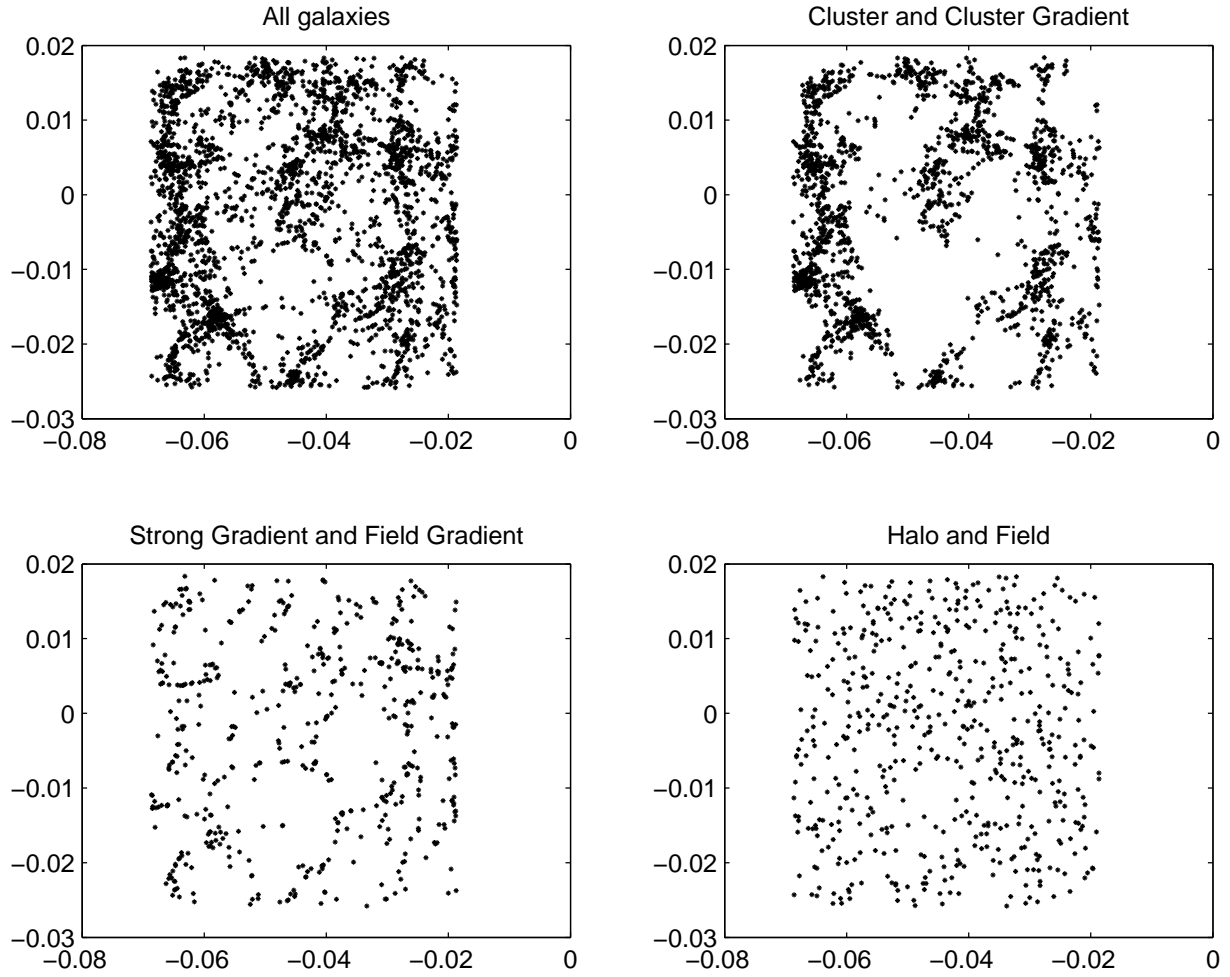


Fig. 7.— Location, in the SOM phase space, of types of galaxies identified by the SOM algorithm: upper left = all galaxies; upper right = *Cluster* and *Cluster Gradient* classes; lower left = *Strong Gradient* and *Field Gradient* classes; lower right = *Halo* and *Field* classes.

For the classes listed in Table 2 Figure 8 presents scatter plots of the input parameters ($R_{Voronoi}/d_{uniform}$ vs $d_{CM}/R_{Voronoi}$) of the SDSS, Millennium Simulation (MS), and our uniform synthetic data, along with class boundaries. The SDSS and MS data are similar, but the MS data spans a slightly larger range of gradients, $d_{CM}/R_{Voronoi}$. There are also subtle but significant differences in the class structure. While the SDSS and MS data sets both contained the same classes, the *Halo* and *Field* classes in the MS data contained more samples and occupied significantly larger regions in phase space, while the three *Gradient* classes were correspondingly smaller.

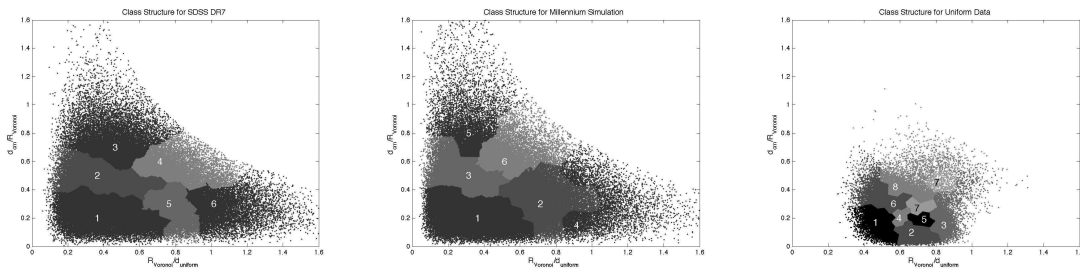


Fig. 8.— Locations, in the neighbor-distance/cell-volume space, of the galaxies assigned to the various SOM classes. Left panel: SDSS DR7 data; middle panel: Millennium Simulation data; right panel: spatially uniform random distribution.

The class structure of the uniform data is noticeably different. Even though the number of samples was similar, they occupy a much smaller region in phase space, with a significantly smaller range of densities and much fewer samples with large gradients. The distribution is sufficiently uniform that the SOM/Bayesian Block technique does not identify any stable classes, and places class boundaries at arbitrary locations. The figure shows a typical result from among the bagged samples, with a large number of poorly-defined classes that in no way resemble the well-ordered structure observed with the SDSS and MS data.

6. Results

Comparison of the results of the three methods, for each of the three data sets, is not entirely straightforward. We have identified a few simple measures to quantify the differences. A future paper will present more detailed quantitative comparisons. In a nutshell, a description of the results of the three methods gives insight into (1) the similarities of the SDSS and Millennium Simulation data sets, (2) the stark differences between them and the uniform distribution regardless of the structure analysis method, and (3) the similarities between the SOM and BB methods, and their differences from the KDE method.

6.1. Classes: From Clusters to the Field

As discussed in §5.3 and demonstrated in Figure 6 the Bayesian Block method yields a series of density levels. Each level contains one or more *blocks*, defined as connected sets of cells each of which is disconnected from all other blocks in the level. The galaxy density within a block is close to the density characterizing the level as a whole, differing only via statistical fluctuations. Obviously blocks correspond directly to structural elements of various densities: blocks of highest density are found in cores of dense clusters, lowest in voids or around isolated field galaxies. Blocks between these extremes trace the intermediate structures of the Cosmic Web. But since the multi-scale structure of the galaxy distribution is characterized by quantities other than local density, blocks do not necessarily correspond directly to physically meaningful structural classes. For example our way of applying Self-organizing maps (§5.4) incorporates density gradient information to generate a set of discrete structural classes (see Figures 7 and 8) which may be more physically significant because their definitions are based on more information than just density. Similarly kernel density estimation incorporates non-local density information by virtue of adaptive smoothing.

Figure 9 depicts how the galaxies are distributed among various classes, one row for each of the three data sets. The histograms in the first column display the distribution of galaxies among the SOM-based classes listed in Table 2. The other columns display, for the other two analysis methods, the distribution of galaxies based solely on their estimated densities in bins chosen to approximately match the resolution of the histograms in the first column, in a way that will now be described.

In this paper we compare the results of the three analysis methods only for galaxies in the highest density classes. This is because they contain the most easily identifiable structures – readily identified with clusters of galaxies. More complete comparisons will be presented in a later paper. Because there is neither a one-to-one or strictly monotonic relation between the density classes uncovered by the three analysis methods we adopted the following procedure. For each of the two non-SOM methods (BB and KDE), start from the high density end and include the maximum number of the corresponding classes¹⁰ such that the total number of galaxies included does not exceed the number of galaxies in the SOM *Cluster* class (ID number 1 in Table 2). For example, in Table 4 one sees that the SOM *Cluster* class contains 44336 galaxies in the SDSS dataset. To reach a similar number of galaxies in the BB method we utilize BB classes 1-4, which sum to 38293 galaxies (see Table 3). Similarly KDE classes 1-4 contain 18286 galaxies.

¹⁰*I.e.* the density levels described at the end of §5.1 for KDE and in §5.3.1 for BB.

Table 3: Number of galaxies and classes in the SOM *Cluster* class for the each dataset (SDSS, MS, Uniform) and algorithm (SOM, BB, KDE). The third row gives the corresponding percentage of the total volume.

	SDSS			Millennium Simulation			Uniform		
	SOM	BB	KDE	SOM	BB	KDE	SOM	BB	KDE
Number	44336	38293	18286	60945	43645	40500	20008	10017	13279
Classes	1	1-4	1-4	1	1-3	1-6	1	1-4	1-2
Volume	12%	6%	9%	16%	6%	13%	8%	7%	55%

Table 4: Number of objects in each class for each dataset (SDSS,MS,Uniform) and algorithm (SOM,BB,KDE)

Class	SDSS			Millennium Simulation			Uniform		
	SOM ^a	BB	KDE	SOM	BB	KDE	SOM	BB	KDE
1	44336	166	30	60945	323	81	20008	288	33
2	36689	1038	243	33075	24496	904	31250	1214	13246
3	15367	22724	3318	6968	18826	2478	7801	3134	74819
4	12223	14365	14695	12089	18016	4650	19279	5381	34754
5	16132	15038	33357	30674	17437	10298	17181	11848	6883
6	9244	16738	42548	5176	13353	22089	19424	60437	1777
7		11380	29116		10677	35988	10292	17097	275
8		11436	9748		9211	37847	6597	10692	39
9		10304	916		7410	23726		6546	5
10		7725	19		7877	8634		6991	1
11		6551	1		6296	1924		3215	
12		3968			4771	280		2070	
13		4800			4356	28		1596	
14		3358			2220			940	
15		1890			1689			360	
16		1583			1039			23	
17		645			581				
18		236			312				
19		46			36				

^aSee Table 2 and §5.4 for a description of the SOM classes.

Figures 10, 11, and 12 compare the distributions of densities, indirectly via histograms of Voronoi volumes, for the SDSS, MS, and uniform random data, respectively. Each figure plots three histograms, of the Voronoi volumes of those galaxies in the SOM selected *Cluster* class, and in the counterpart selections for the BB and KDE methods as just defined. The independent variable of these histograms is a common logarithmic binning of the range of Voronoi volumes (labeled with bin number, not to be confused with a class identifier). Even though the KDE method does not use the Voronoi volumes in its calculation of density we rely on the Voronoi volume associated with a given galaxy for all three methods to make the volumes more comparable. The legends for each of these three figures gives the percentage of total cluster volume versus the full volume for each dataset. These numbers also appear in Table 3.

In Figure 10 the easiest distribution to understand is that for the Bayesian Block method. Since it uses the cell-based volumes, solely and directly, the distribution is naturally a broad lump of small (high density) cells, with no tail of larger (low density) ones. In other words its levels are defined directly in terms of the volumes, as depicted in Fig 6. Both of the other methods blend in other non-local information – the SOM explicitly through density gradients, and KDE implicitly via its adaptive kernel – leading to the rather long tails to the high end of the volume distributions. The KDE distribution resides nearly midway between the SOM and BB ones, presumably because of its implicit blend of local and non-local information. Nearly the same pattern as seen in the SDSS is repeated for the Millennium Simulation dataset in Figure 11 for each of the methods and the cluster volume percentages. However, for the Uniform dataset in Figure 12 the SOM and BB cluster classes appear very similar in volume percentage, while the corresponding KDE classes contain many more galaxies.

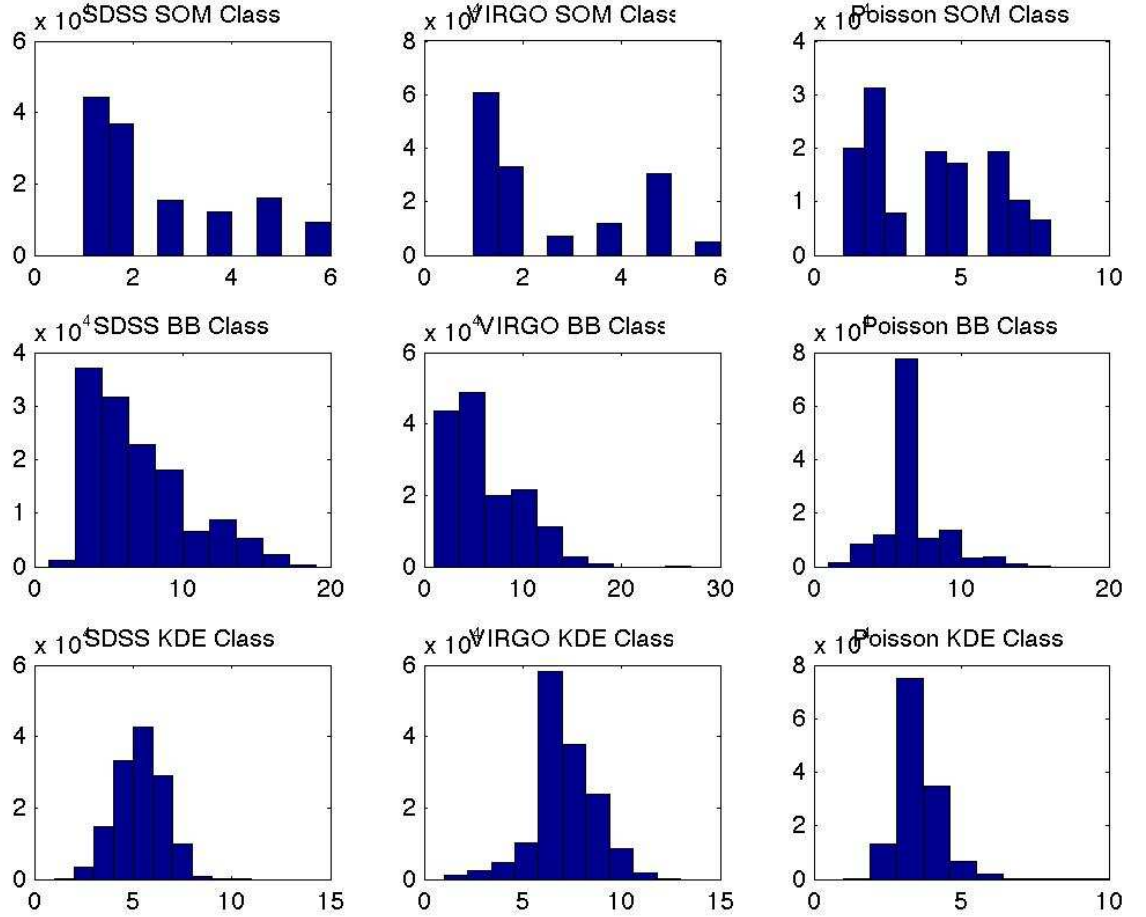


Fig. 9.— These histograms show the number of points in each class, for the three methods applied to the three data sets. The columns indicate the analysis method: 1 = SOM: Self-organizing map; 2 = BB: Bayesian block; 3 = KDE: Kernel density estimator. In first column the bins are the natural classed yielded by the SOM; the other two are approximately matched density bins, as described in the text. The rows indicate the data analyzed: 1 = SDSS; 2 = MS: Millennium Simulation; 3 = spatially uniform random distribution.

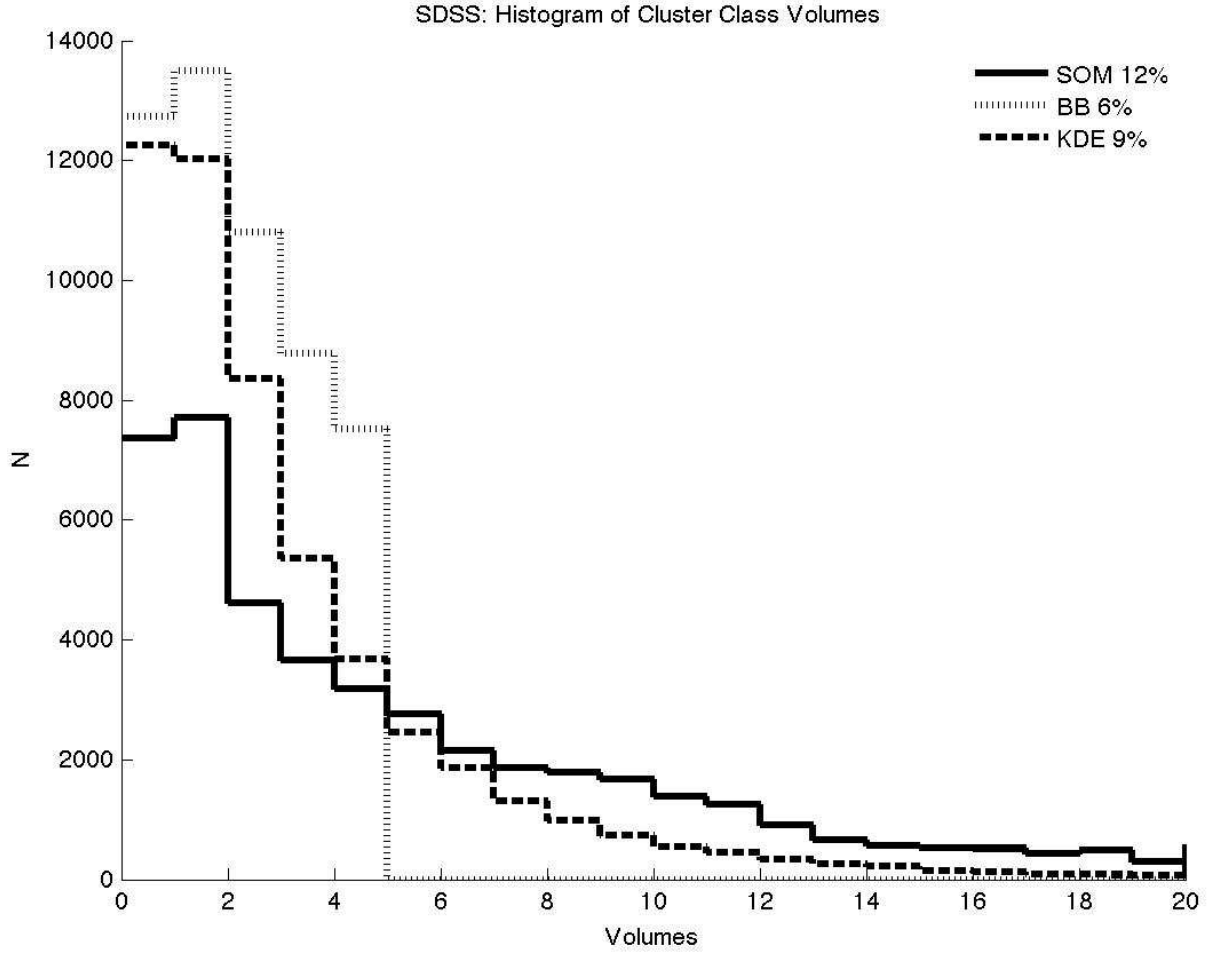


Fig. 10.— Volume distributions for the SDSS cluster classes, in equal logarithmic bins. The legend describes the percentage of *Cluster* class Voronoi volumes for each method.

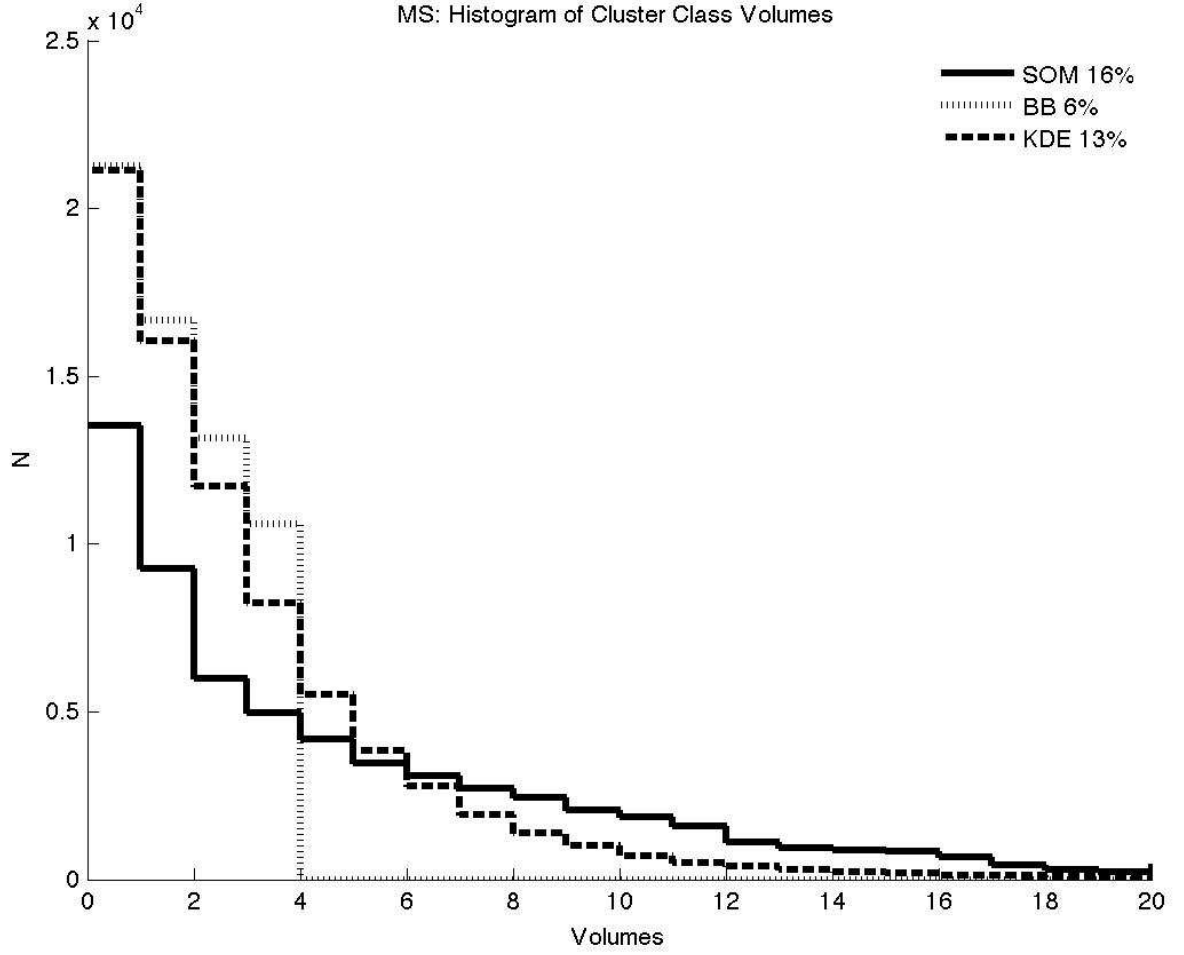


Fig. 11.— Volume histograms for the Millennium Simulation cluster classes. The legend describes the percentage of *Cluster* class Voronoi volumes for each method.

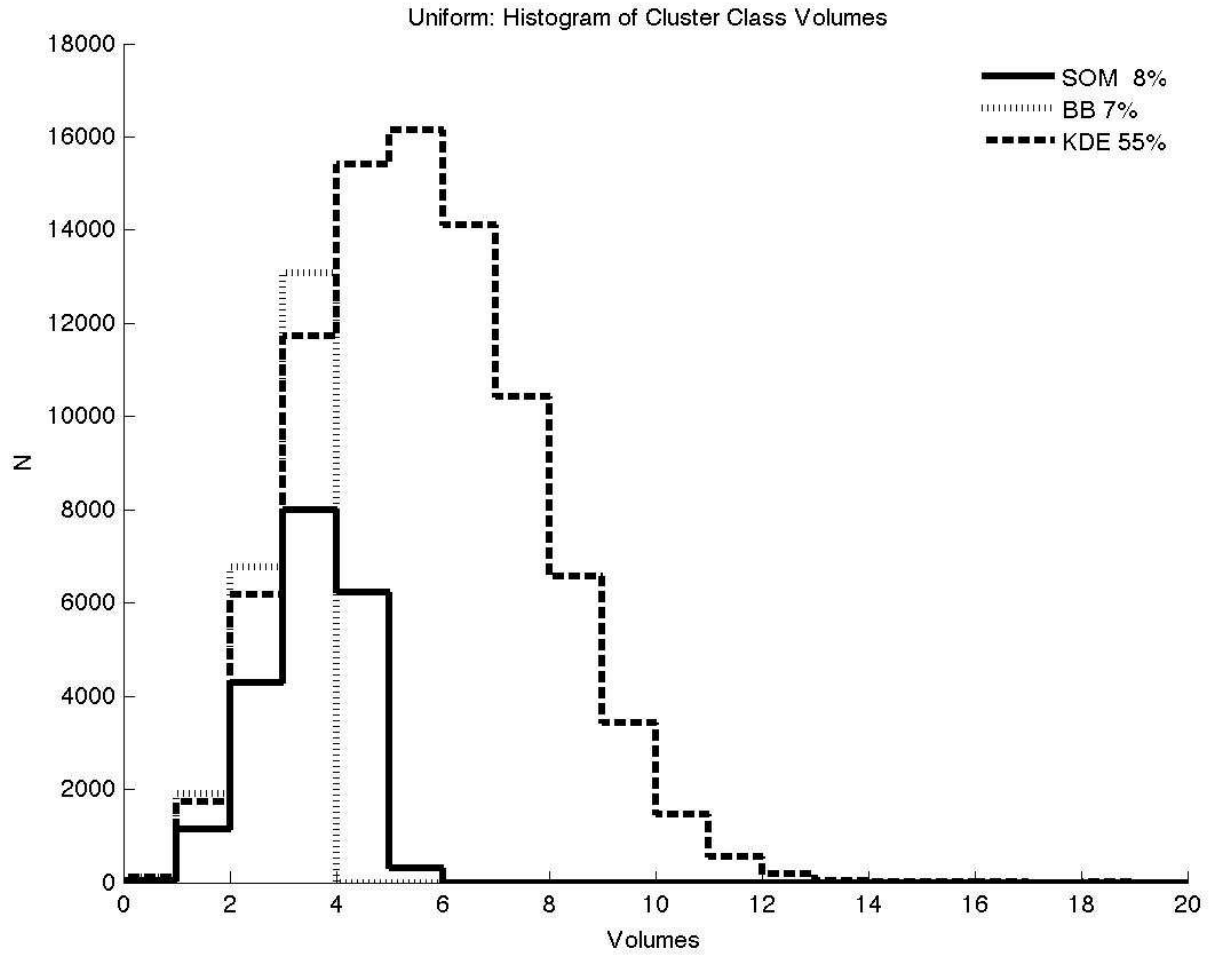


Fig. 12.— Volume histograms for the uniform cluster classes. The legend describes the percentage of *Cluster* class Voronoi volumes for each method.

6.2. Visualizing high density classes

A thin spatial slice (from a fixed viewing angle) of the galaxies found in the high density classes just described in §6.1, for each method and dataset are compared side-by-side in Figure 13. This figure collects the view shown in the central panels of the 3×3 plots from Figures 14-16, 20-22 and 26-28, below. The three methods identify similar structures in the SDSS and MS data, but of course not in the uniformly random data. In the bottom row note that the three methods select markedly different depths of the upper end of the density distribution (*cf.* the right-hand panel of Figure 3) but do not falsely reveal medium or large scale structure.

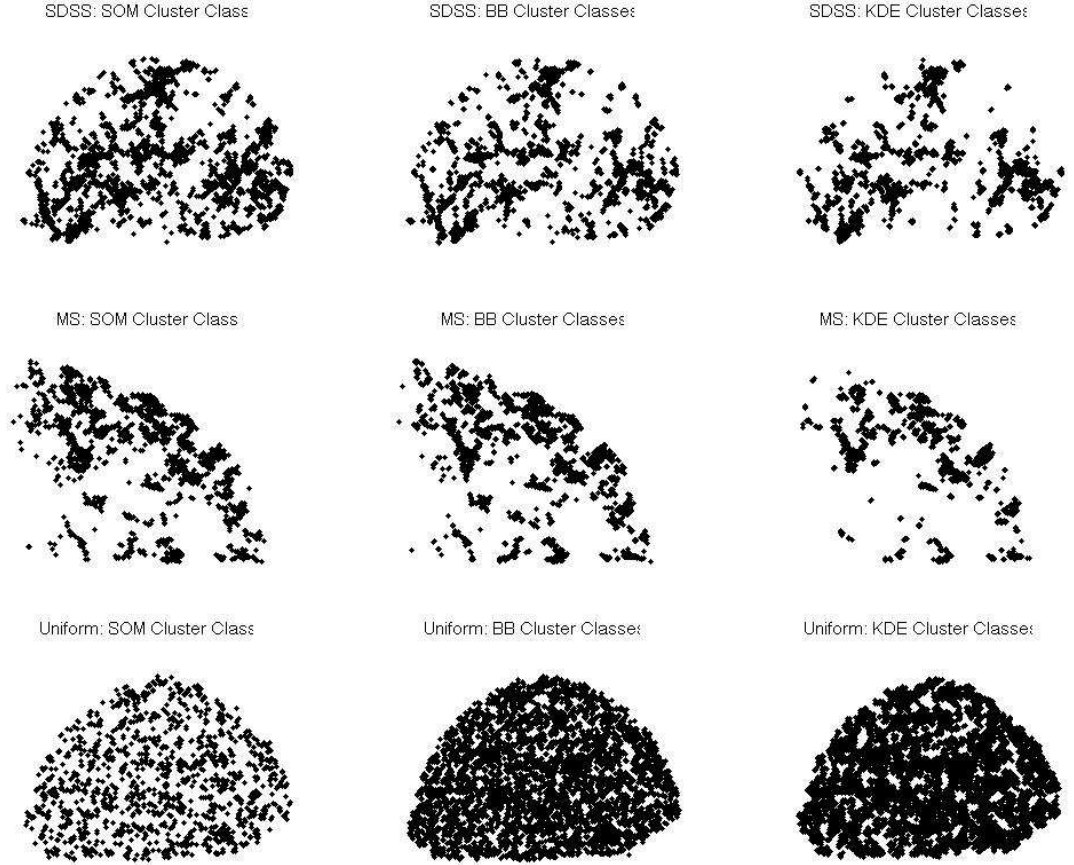


Fig. 13.— Projections of the spatial locations of the main density structures found with the three methods, using the three data sets. These are the central plot from Figures 14 – 16, 20 – 22 and 26 – 28. As discussed in the text in some sense these structures are clusters, but they are defined simply as localized density peaks. From left to right: Bayesian blocks, SOM clusters, and KDE peaks. Top to bottom: spatially uniform random distribution, SDSS DR7, and Millennium Simulation.

The remaining figures of this section elucidate clustering associated with the highest density regions for the three analysis methods, with sets of figures for the SDSS, the MS, and the uniformly random data. Begin with three spatial distributions for the SDSS data, Figures 14-16, as derived with SOM, BB, and KDE respectively. The rows in figure 14 show

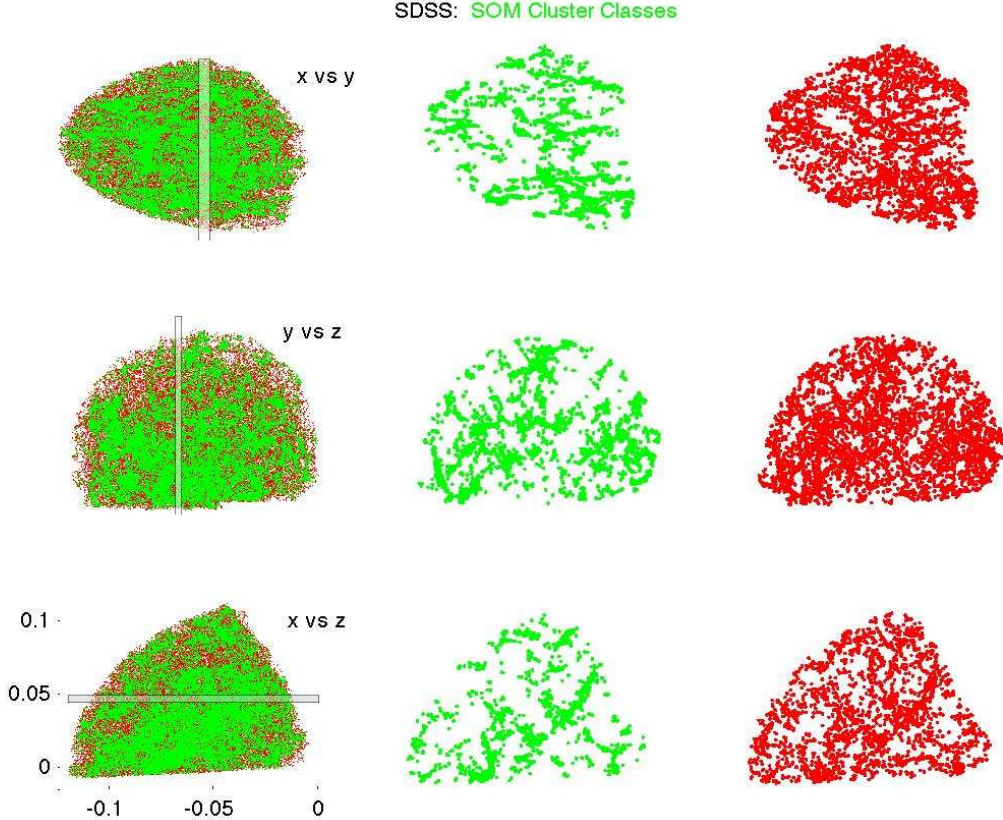


Fig. 14.— Self-organizing map analysis of the Volume Limited SDSS data. The three rows in each column show the locations of the derived block structures in three orthogonal projections. Column 1: The green points (found in the SOM *Cluster* class) are those assigned higher densities by the SOM algorithm, while the red are all other points. For clarity the corresponding points in thin spatial slices (indicated as gray bands in Column 1) are plotted in green (points in the SOM *Cluster* class) and red (non-cluster points) in Columns 2 and 3 respectively.

SOM-derived structures in three orthogonal projections, the first column being the entire data-cube (see Figure 1 and §4). The green points are galaxies in the SOM *Cluster* class, while red points are not. The remaining two columns differ from the first in two ways: they show only galaxies within thin spatial slices (delineated as light gray bands in Column 1), and they separate the cluster and non-cluster galaxies (displayed in gray and black, respectively,

in all 3 columns) to better reveal the structures and the gross differences in the distributions of *Cluster* galaxies and non-cluster galaxies.

Figure 15 presents the same display pattern for the BB analysis, and Figure 16 for the KDE analysis. The SOM and BB cluster classes appear to be relatively similar, while the KDE appears markedly different from the other two, although some structures do appear more or less the same with all three analysis algorithms.

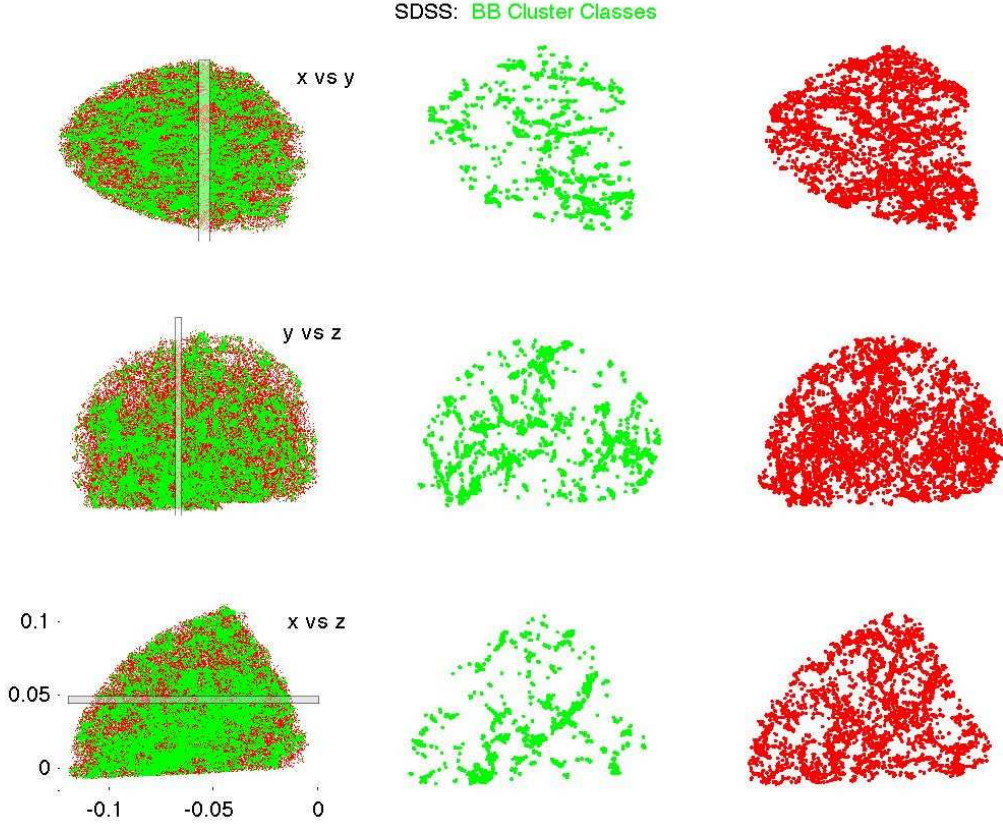


Fig. 15.— The same as Figure 14, but for the Bayesian Block (BB) Structure analysis of the Volume Limited SDSS data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the BB algorithm, while the red are all other points. Many of these green points would be considered to be in high-density clusters and are what we consider to constitute the BB *Cluster* class. Column 2 shows the same BB structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

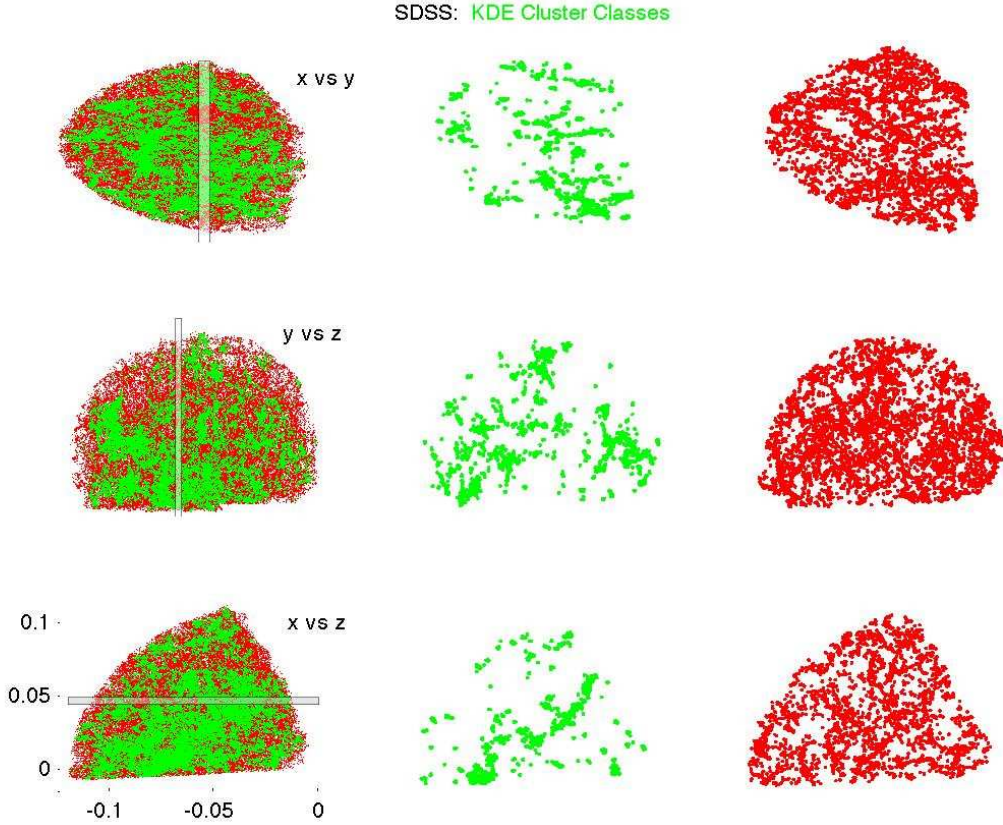


Fig. 16.— The same as Figure 14, but for the Kernel Density Estimation (KDE) analysis of the Volume Limited SDSS data. The three rows in each column show the locations of the KDE derived structures in three different projections. Column 1: The green points are considered to be in high-density clusters and are what we consider to constitute the KDE *Cluster* class, while the red are all other points. Column 2 shows the same KDE structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

Continuing the discussion of the SDSS, now turn to a somewhat detailed look at the distribution of the galaxies over various classes that have been defined above. The next three plots, Figures 17, 18, and 19, show histograms of the classes for the three methods applied to the SDSS dataset. Figure 17 plots the BB classes on the x-axis and the KDE ones on the y-axis. The number of KDE objects in a given BB class for a given KDE class is shown in the corresponding histogram bin.

Ignoring the coloring scheme for the moment, in Figure 17 one sees a clear correlation between the density classes (indicated inversely by the class number labels on the axes) in the KDE and BB classifications. To wit, KDE class 1 through 4 objects (see Table 3) are found exclusively in BB classes 1 through 7 – implying that there are no KDE-class 1 through 4 objects in BB classes 8 through 19. The coloring scheme used for the individual histograms is intended to show how the method not plotted on either the x or y axis distributes its cluster classes in green in the other two method classes. Non-cluster classes are in red. For example, for this Figure 17 the method not plotted on the x (BB) or y (KDE) axes is the SOM method. The SOM cluster class is plotted in green and all other SOM classes are in red. Most of the SOM cluster class objects show up in KDE classes 2–7 with a few in class 8. All of the SOM cluster class objects appear in BB classes 1–10. None of the SOM cluster class objects show up in the lowest density BB classes 11–19 or KDE class 9. Clearly the overlap between the cluster classes of one method and the non-cluster classes of others is not insignificant, in accordance with the fact that the structural classifications carried out by the three methods are based on different information content.

Figures 18 and 19 are identical to 17, but for the other two combinations of variables assigned to the x- and y- axes (in both cases including the third variable with the shown histograms).

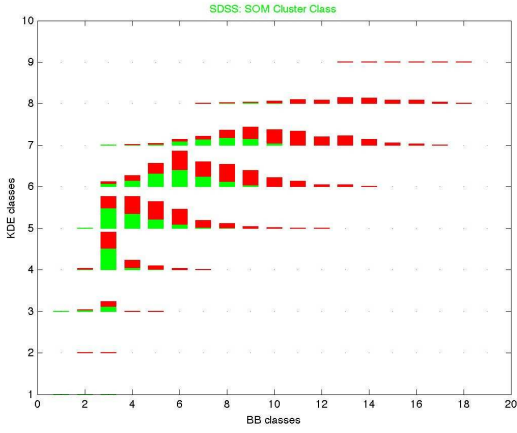


Fig. 17.— For the SDSS data, this figure compares high and low-density classes from the 3 methods. Each of the 9 sets of histograms shows the distribution among the BB classes (horizontal axis) of those in the corresponding KDE classes (indicated on the vertical axis). The full distribution over the SOM clusters is not shown, but in each histogram bar the SOM defined *Cluster* class is in green. The SOM non-cluster classes are in red.

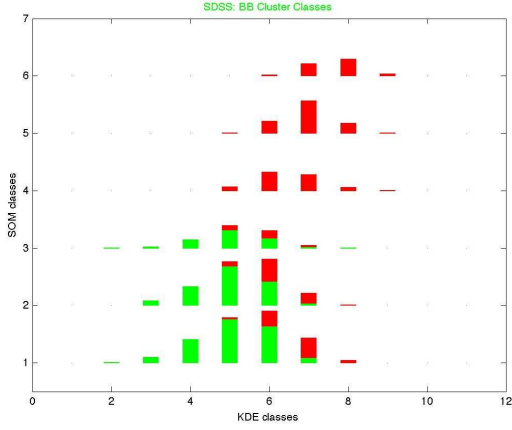


Fig. 18.— Also for SDSS data, and similar to Figure 17, this figure compares the high and low-density classes from the 3 methods. Each of the 6 histograms shows the distribution among the KDE classes (horizontal axis) of those in the corresponding SOM classes (indicated on the vertical axis). The full distribution over the BB classes is not shown, but in each histogram bar the high-density BB *Cluster* classes are in green. Non-high-density BB classes are in red.

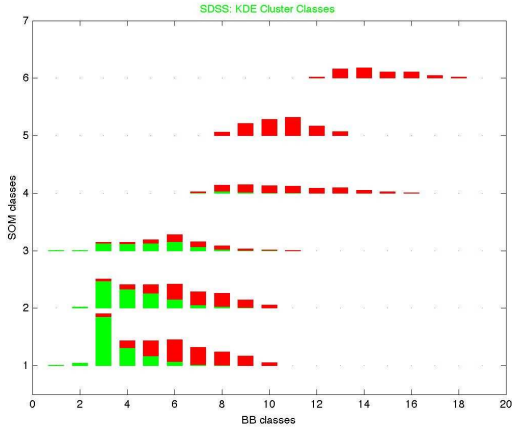


Fig. 19.— Also for the SDSS data, and similar to Figure 17, this figure compares the high and low-density classes from the 3 methods. Each of the 6 histograms shows the distribution among the BB classes (horizontal axis) of those in the corresponding SOM classes (indicated on the vertical axis). The full distribution over the KDE classes is not shown, but in each histogram bar the high-density KDE *Cluster* classes are in green. Non-high-density KDE classes are in red.

Having discussed the example results for the actual SDSS data, we now present an exactly parallel set of figures for the artificial data contained in the Millennium Simulation data, as described in §4. The first three spatial distribution plots for MS, Figures 20 – 22, are parallel to Figures 14 – 16, discussed above for the SDSS data. These are followed by the class distribution plots in Figures 23 – 25, parallel to those in Figures 17 – 19.

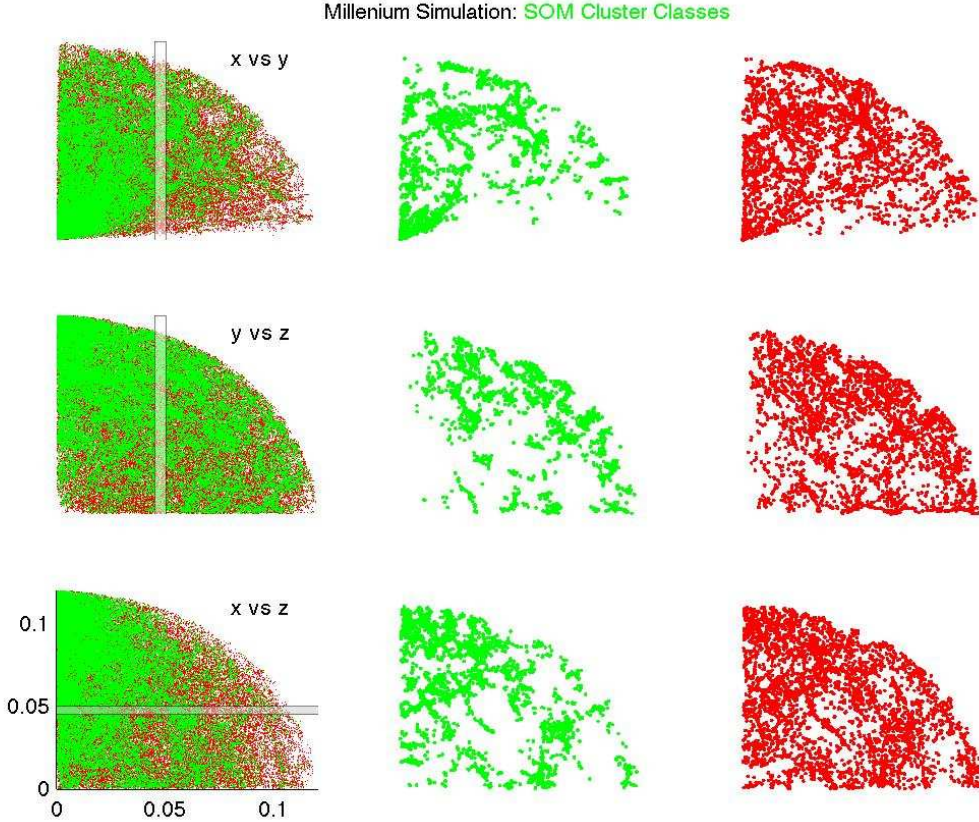


Fig. 20.— Similar to Figure 14, but instead the Self-organizing map (SOM) analysis of the Volume Limited Millennium Simulation (MS) data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the SOM algorithm (found in the *SOM Cluster* class), while the red are all other points. Column 2 shows the same *SOM Cluster* structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

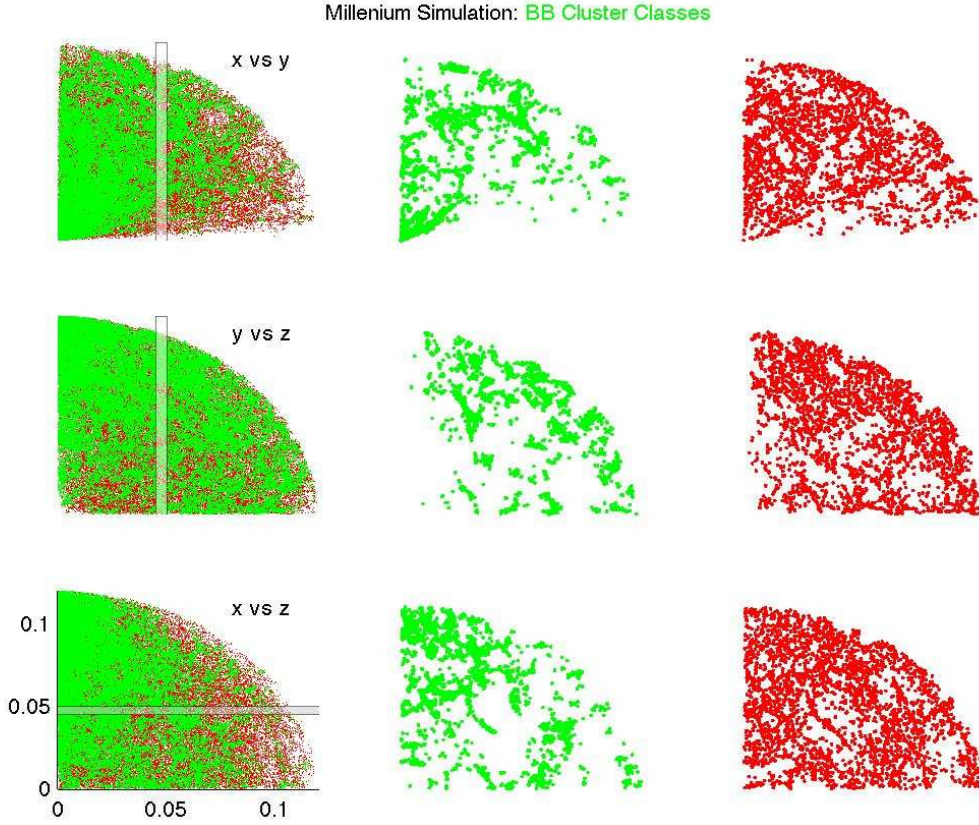


Fig. 21.— The same as Figure 20, but for the Bayesian Block (BB) Structure analysis of the Volume Limited MS data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the BB algorithm (found in the BB *Cluster* class), while the red are all other points. Column 2 shows the same BB *Cluster* structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

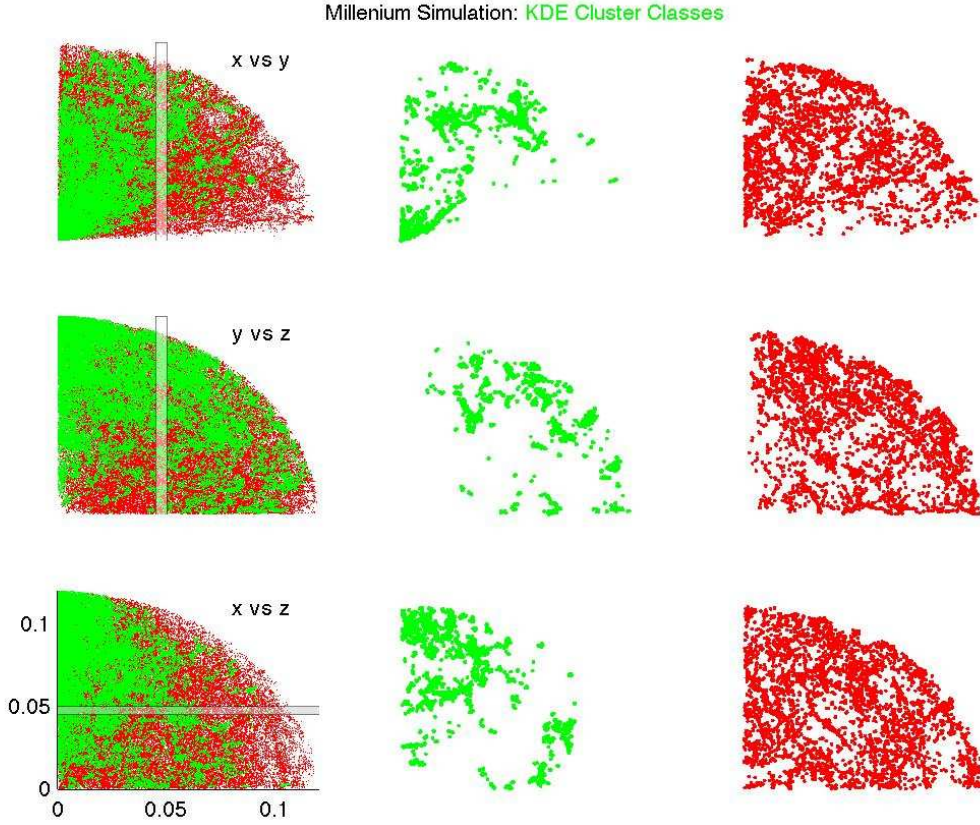


Fig. 22.— The same as Figure 20, but for the Kernel Density Estimation (KDE) analysis of the Volume Limited MS data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the KDE algorithm (found in the KDE *Cluster* class), while the red are all other points. Column 2 shows the same KDE *Cluster* structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

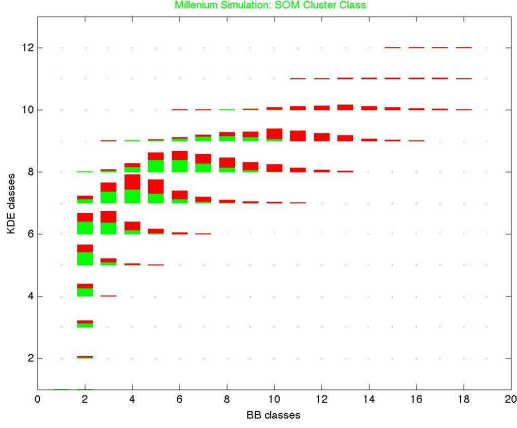


Fig. 23.— For the Millennium Simulation (MS) data, this figure compares high and low-density classes from the 3 methods. Each of the 12 sets of histograms shows the distribution among the BB classes (horizontal axis) of those in the corresponding KDE class (indicated on the vertical axis). The full distribution over the SOM classes is not shown, but in each histogram bar the SOM defined *Cluster* class is in green. The SOM non-cluster classes are in red.

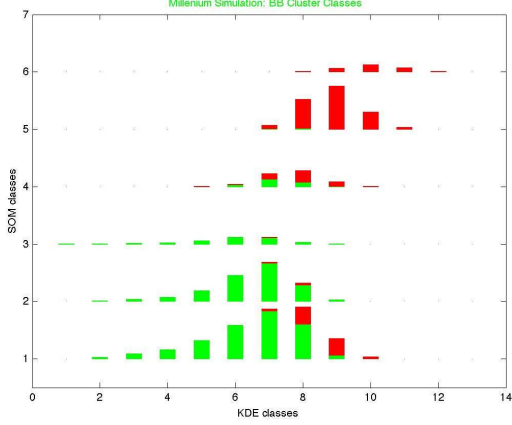


Fig. 24.— Also for MS data, and similar to Figure 23. This figure compares high and low-density classes from the 3 methods. Each of the 6 sets of histograms shows the distribution among the KDE classes (horizontal axis) of those in the corresponding SOM class (indicated on the vertical axis). The full distribution over the BB classes is not shown, but in each histogram bar the BB defined *Cluster* classes are in green. The BB non-cluster classes are in red.

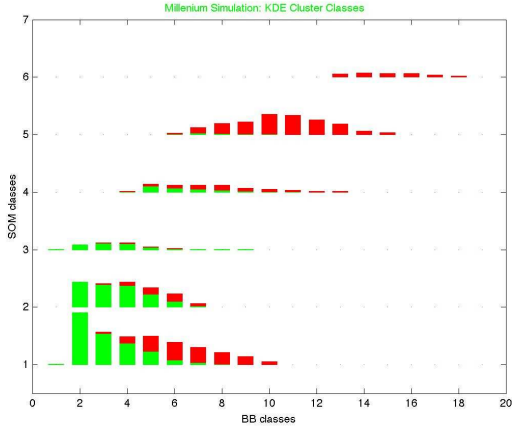


Fig. 25.— Also for MS data, and similar to Figure 23. This figure compares high and low-density classes from the 3 methods. Each of the 6 sets of histograms shows the distribution among the BB classes (horizontal axis) of those in the corresponding SOM class (indicated on the vertical axis). The full distribution over the KDE classes is not shown, but in each histogram bar the KDE defined *Cluster* classes are in green. The KDE non-cluster classes are in red.

Having discussed the example results for the actual SDSS data, and the Millennium Simulation data, we now present an exactly parallel set of figures for the artificial data contained in the uniformly and randomly distributed data, as described in §4.

The first three spatial distribution plots for the uniformly random data Figures 26 – 28 are parallel to Figures 14 – 16 discussed above for the SDSS data, and Figures 20 – 22 discussed above for the MS data. These are followed by the class distribution plots in Figures 29 – 31, parallel to those in Figures 23 – 25, and Figures 17 – 19.

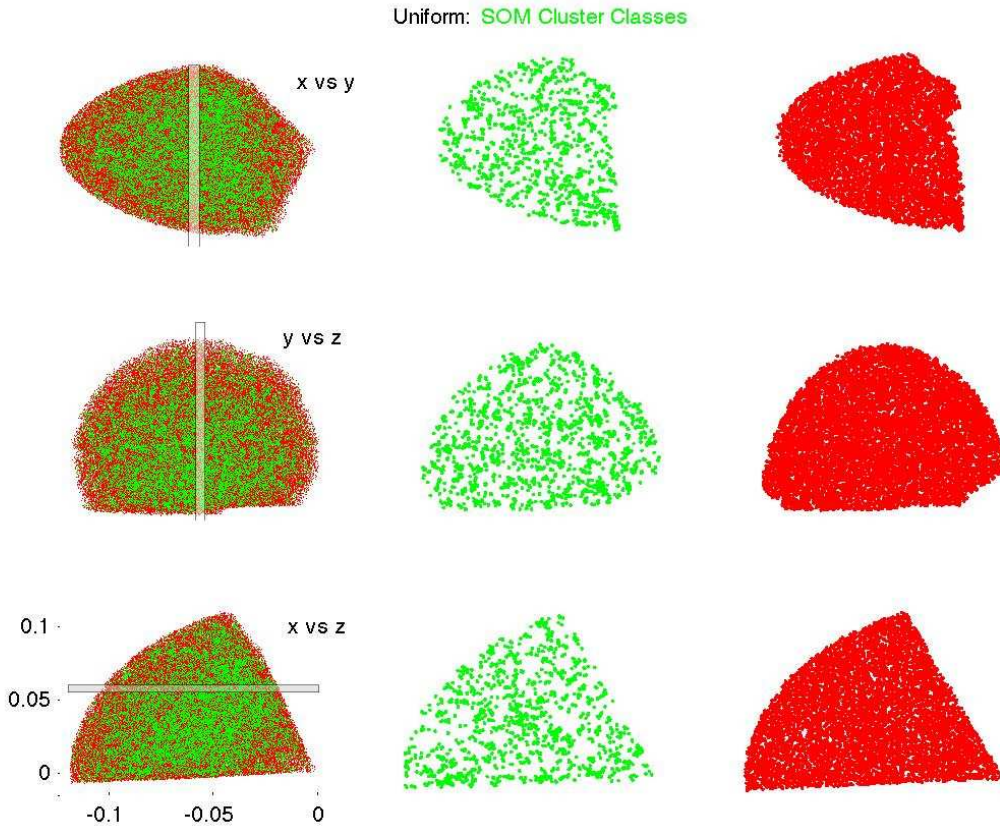


Fig. 26.— Self-organizing map (SOM) analysis of the spatially uniform random distribution data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the SOM algorithm (found in the *SOM Cluster* class), while the red are all other points. Column 2 shows the same SOM structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

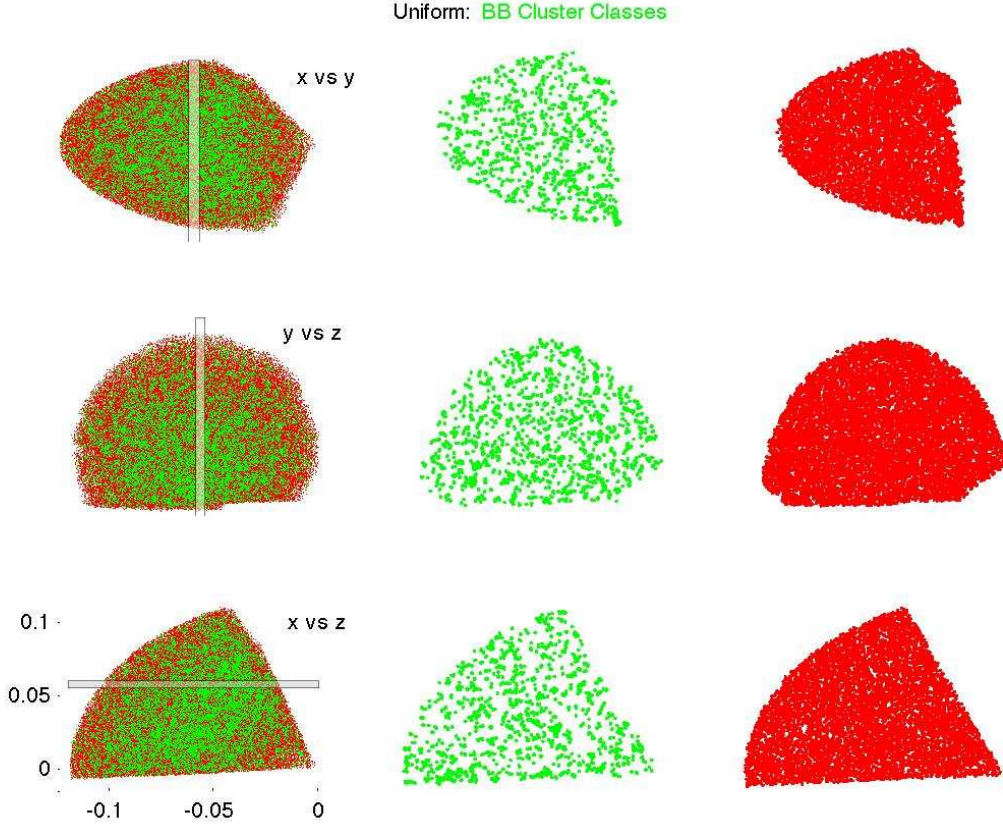


Fig. 27.— The same as Figure 26, but for the Bayesian Block (BB) analysis of the spatially uniform random distribution data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the BB algorithm (found in the BB *Cluster* class), while the red are all other points. Column 2 shows the same BB structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

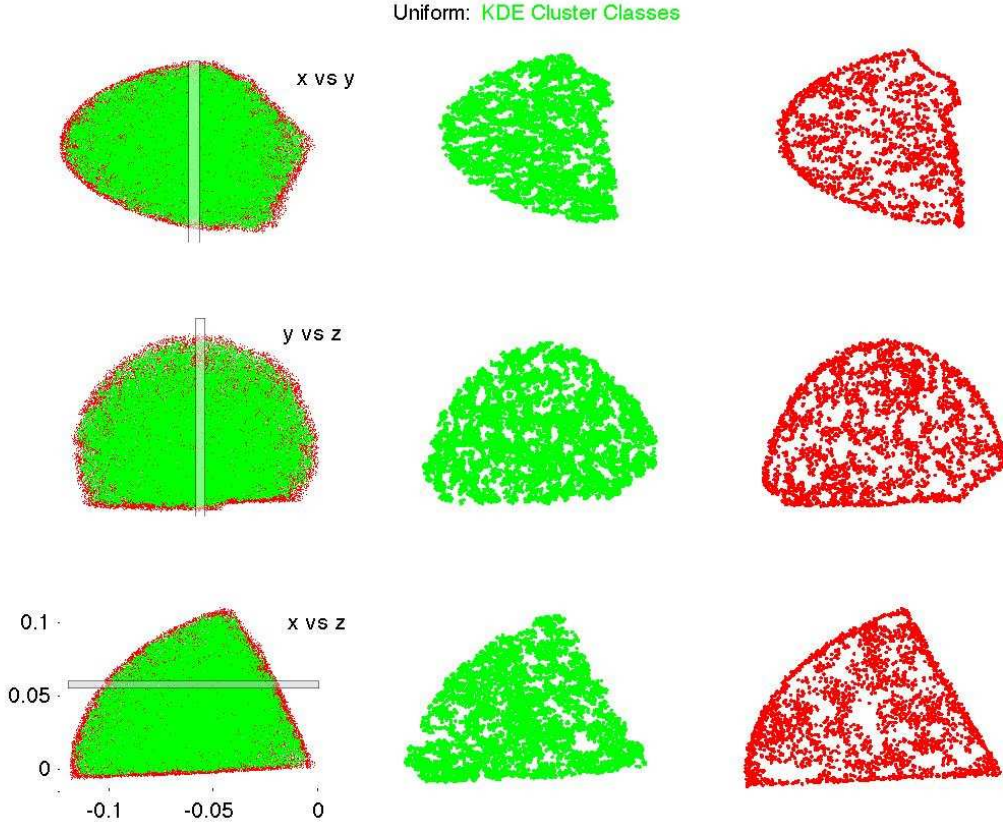


Fig. 28.— The same as Figure 26, but for the Kernel Density Estimation (KDE) analysis of the spatially uniform random distribution data. The three rows in each column show the locations of the derived block structures in three different projections. Column 1: The green points are those assigned higher densities by the KDE algorithm (found in the KDE *Cluster* class), while the red are all other points. Column 2 shows the same KDE high density structures in a thin slice, to better visualize these results. Column 3 is the complement of column 2: all structures not selected in the same thin slice shown in column 2.

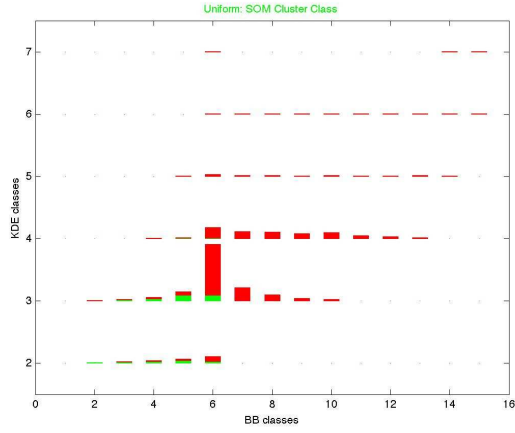


Fig. 29.— For the spatially uniform random distribution data, this figure compares high and low-density classes from the 3 methods. Each of the 6 sets of histograms shows the distribution among the BB classes (horizontal axis) of those in the corresponding KDE class (indicated on the vertical axis). The full distribution over the SOM classes is not shown, but in each histogram bar the SOM defined *Cluster* class is in green. The SOM non-cluster classes are in red.

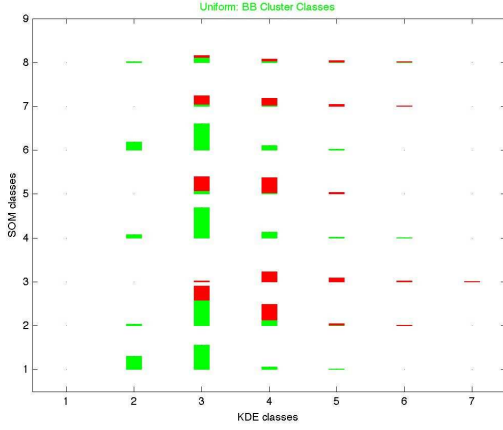


Fig. 30.— Also for spatially uniform random distribution data, and similar to Figure 29, this figure compares high and low-density classes from the 3 methods. Each of the 8 sets of histograms shows the distribution among the SOM classes (horizontal axis) of those in the corresponding KDE class (indicated on the vertical axis). The full distribution over the BB classes is not shown, but in each histogram bar the BB defined *Cluster* classes are in green. The BB non-cluster classes are in red.

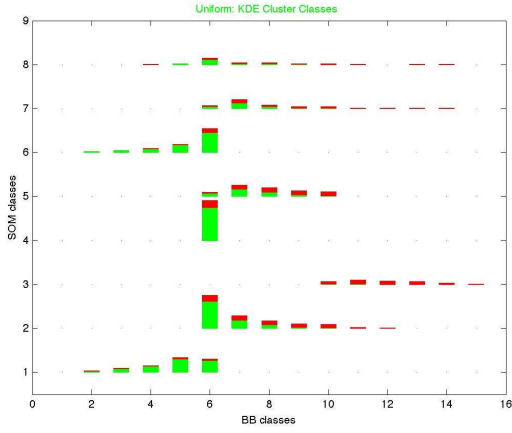


Fig. 31.— Also for the spatially uniform random distribution data, and similar to Figure 29, this figure compares high and low-density classes from the 3 methods. Each of the 8 sets of histograms shows the distribution among the BB classes (horizontal axis) of those in the corresponding SOM class (indicated on the vertical axis). The full distribution over the KDE classes is not shown, but in each histogram bar the KDE defined *Cluster* classes are in green. The KDE non-cluster classes are in red.

In all cases there is very little evidence of clustering in the uniformly distributed points, exactly as one would expect. The average densities are again very similar for the BB and SOM. The KDE appears to select more galaxies for its cluster class, while explicitly avoiding the majority of galaxies at the border; this odd behavior was not demonstrated in the other datasets, but is likely just an edge-effect that could easily be removed.

Table 5 distills the cluster class overlap between methods into a single table as shown in Figures 17 – 19, 23 – 25, and 29 – 31. For the most part these summaries for the SDSS and MS cases are more alike than not, whereas those for the uniformly random case are very different. It is clear that all three algorithms assign high density regions to classes in somewhat different ways, just as one would expect.

Table 5: This table addresses how much the different methods assign galaxies (at the high density end of the distribution) to the same/different classes. Entries indicate which classes (defined by the method labeled in the second row from the top of the columns, and for the data set indicated in the first row) are contained in the cluster class for the method indicated in the left-most column.

	SDSS			Millennium Simulation			Uniform		
	SOM	BB	KDE	SOM	BB	KDE	SOM	BB	KDE
SOM <i>Cluster</i> class	1	1–6	2–7	1	1–10	2–10	1	-	-
BB <i>Cluster</i> classes	1–3	1–4	2–7 ^a	1–3	1–3	2–9	1,6	1–4	2–4
KDE <i>Cluster</i> classes	1–3	1–7	1–4	1–4	1–7	1–6	2–6	2–6	1–2

^aFor example, how many KDE classes are found in the BB cluster classes (1–4)? In this case KDE classes 2–7 contain BB cluster classes 1–4.

7. Summary and Conclusions

We have described two techniques newly applied to characterize structures in large 3-D galaxy surveys based on Voronoi tessellation – “Bayesian Blocks” (BB) and “Self-organizing maps” (SOM). These two new techniques were compared with a third well known technique called Kernel Density Estimation (KDE).

The techniques were applied to three example datasets. The first was a volume limited sub-sample of the SDSS Data Release 7. The second was a volume limited sub-sample of the Millennium Simulation, while the 3rd was a uniform randomized set of points similar in size to the other two. The BB and SOM methods proved to pick similar high-density structures from the SDSS and Millennium Simulation datasets. The KDE method generally gives rather different results, although it was able to identify some of the same high-density structures. The uniform randomized sample proved a challenge to all three techniques ability to discern statistically significant high-density concentrations – as it should have, since they don’t exist.

In future publications we plan to provide more details on the analysis previewed here, including preparation of an all-scale structure catalog (distinguishing from the term *large-scale structure*). Our catalog will include features unique to our analysis approach, such as:

- internal comparison between structures which have been found using two different analysis methods, but which can be reliably identified as comprising the same physical structure, say based on spatial coincidence.
- measures of convexity/concavity and their distributions
- the sizes and directions of tri-axial ellipsoids fit to the blocks,
- other morphological quantities

This will allow us to further compare our self-organizing map and Bayesian block analysis on the Sloan Digital Sky Survey data with other workers’ results including catalogs of clusters, sheets (walls), filaments, voids, *etc.*

Certainly the reader may be skeptical of any one of the three methods abilities to distinguish between similar structures in SDSS redshift data such as Fingers-of-God and line-of-sight filaments. However, given our ability to obtain the “ground truth” from the original Millennium Simulation positions (x, y, z) and velocities (V_x, V_y, V_z) we believe it will

be possible characterize and distinguish structures that mimic each other in SDSS type data sets.

We are grateful to the NASA-Ames Director’s Discretionary Fund and to Joe Bredekamp and the NASA Applied Information Systems Research Program for support and encouragement. We thank the Institute for Pure and Applied Mathematics at UCLA and the Banff International Research Station for hospitality over times where some of this work was carried out. Helpful discussions and suggestions over the years came from Chris Henze, Creon Levit, and Ashok Srivastava.

Thanks goes to Ani Thakar and Maria Nieto-Santisteban for their help with our many SDSS casjobs queries. Michael Blanton’s help with using his SDSS NYU VAGC catalog were also very much appreciated. Zeljko Ivezic, Robert Lupton, Jim Gray and Alex Szalay also provided essential help in utilizing the SDSS.

Funding for the SDSS has been provided by the Alfred P. Sloan Foundation, the Participating Institutions, the National Aeronautics and Space Administration, the National Science Foundation, the U.S. Department of Energy, the Japanese Monbukagakusho, and the Max Planck Society. The SDSS Web site is <http://www.sdss.org/>.

The SDSS is managed by the Astrophysical Research Consortium for the Participating Institutions. The Participating Institutions are The University of Chicago, Fermilab, the Institute for Advanced Study, the Japan Participation Group, The Johns Hopkins University, Los Alamos National Laboratory, the Max-Planck-Institute for Astronomy, the Max-Planck-Institute for Astrophysics, New Mexico State University, University of Pittsburgh, Princeton University, the United States Naval Observatory, and the University of Washington.

This research has made use of NASA’s Astrophysics Data System Bibliographic Services.

This research has also utilized the viewpoints (Gazis, Levit, & Way 2010) software package.

A. Appendix: SDSS casjobs query

```
Select p.ObjID, p.ra, p.dec,  
p.dered_u, p.dered_g, p.dered_r, p.dered_i, p.dered_z,  
p.Err_u, p.Err_g, p.Err_r, p.Err_i, p.Err_z,  
s.z, s.zErr, s.zConf  
FROM SpecOBJall s, PhotoObjall p
```

WHERE s.specobjid=p.specobjid
and s.zConf>0.95 and s.zWarning=0 and
(p.printtarget & 0x00000040 > 0)
and (((flags & 0x8) = 0) and ((flags & 0x2) = 0) and ((flags & 0x40000) = 0))

B. Appendix: Catalog Attributes

Table 6: Attributes

u, g, r, i, z	Apparent magnitudes from the SDSS DR7.
U, G, R, I, Z	Absolute magnitudes from the SDSS DR7.
z, z_{err}	Redshift and the uncertainty in redshift.
$d_{uniform}$	Average spacing between points for a uniform distribution.
d_{1-6}	Distances in units of z to the six nearest neighbors.
$R_{Voronoi}$	$(\text{Voronoi volume})^{1/3}$ in units of z . A measure of local density.
d_{CM}	Distance in z from a galaxy to the CM of its Voronoi cell.
R_{Max}	Maximum distance from the point to a vertex of the Voronoi cell.
R_{Min}	Minimum distance from the point to a vertex of the Voronoi cell.
$R_{Voronoi}/d_{Uniform}$	A dimensionless measure of local density.
$R_{Max}/d_{Uniform}$	A dimensionless measure of R_{Max} .
$R_{Min}/d_{Uniform}$	A dimensionless measure of R_{Min} .
$d_{CM}/R_{Voronoi}$	A dimensionless measure of the local gradient.
‘elongation’	A simple dimensionless measure of the elongation of a Voronoi cell.

REFERENCES

- Abazajian, K.N. et al. 2003, AJ, 126, 2081
- Abazajian, K.N. et al. 2009, ApJS, 182, 543
- Abell, G. O. 1958, ApJS, 3, 211
- Adelman-McCarthy, J. K. et al. 2007, ApJS, 172, 634
- Andersen, P., Borgan, O., Gill, R. & Keiding, N. 1992, *Statistical Models Based on Counting Processes*, Springer-Verlag: New York.
- Aragón-Calvo, M.A., Jones, B.J.T., van de Weygaert, R. & van der Hulst, J.M. 2007, A&A, 474, 315
- Aragón-Calvo, M.A., van de Weygaert, R. & Jones, B.J.T. 2010, MNRAS, 408, 2163
- Aragón-Calvo, M.A., Shandarin, S.F. & Szalay, A. 2010 in Press, “Geometry of the Cosmic Web: Minkowsky Functionals from the Delaunay Tessellation,” ISVD10 (Seventh International Symposium on Voronoi Diagrams in Science and Engineering), Quebec City, Canada. IEEE CPS, ed. M.A. Mostafavi
- Balogh, M. et al. 2004, MNRAS348, 1355
- Barber, C.B., Dobkin, D.P. & Huhdanpaa, H.T. 1996, “The Quickhull algorithm for convex hulls,” ACM Transactions on Mathematical Software, 22(4):469-483, Dec 1996, <http://www.qhull.org>
- Barrow, J.D., Bhavsar, S.P & Sonoda, D.H. 1985, MNRAS, 216, 17
- Bauer, H.-U. & Villmann, T., 1997, “Growing a Hypercubical Output Space in a Self-Organizing Feature Map.” IEEE Transactions on Neural Networks, 8(2):218-226.
- Beaky, M.M., Scherrer, R.J. & Villumsen, J.V. 1992, ApJ, 387, 443
- Benson A.J., Frenk C.S., Baugh C.M., Cole S., LaceyC.G., 2001, MNRAS, 327, 1041
- Blanton, M. R. et al. 2005, AJ129, 2562
- Blanton, M.R., Eisenstein, D., Hogg, D.W. & Zehavi, I. 2006, ApJ, 645, 977
- Blanton, M.R., & Berlind, A.A. 2007, ApJ, 664, 791
- Bok, B. 1934, Havard College Obs. Bull., 895, 1

- Bond, N., Strauss, M.A., & Cen, R. 2009, arXiv:0903.3601v1
- Botzler, C.S., Snigula, J., Bender, R. & Hopp, U. 2004, MNRAS, 349, 425
- Buryak, O.E., Doroshkevich, A.G. & Fong, R. 1994, ApJ, 434, 24
- Butcher, H. & Oemler, A. 1978, ApJ, 226, 559
- Canavezes, A. et al. 1998, MNRAS, 297, 777
- Cappellari, M. 2009, Voronoi binning: Optimal adaptive tessellations of multi-dimensional data **astro-ph:0912.1303** Invited review for the volume “Tessellations in the Sciences: Virtues, Techniques and Applications of Geometric Tilings”, eds. R. van de Weijgaert, G. Vegter, J. Ritzerveld and V. Icke, Kluwer/Springer (submitted).
- Choi, E., Bond, N.A, Strauss, M.A., Coil, A.L, Davis, M. & Willmer, C.N.A. 2010, MNRAS, 406, 320
- Choi, Y., et al. 2010, arXiv:1005.0256v1
- Colberg, J.M. 2007, MNRAS, 375, 337
- Colberg, J.M. et al. 2008, MNRAS, 387, 933
- Coles, P. 1990, Nature, 346, 446
- Colless, M.M. et al. 2001, MNRAS, 328, 1039
- Connolly, A.J. et al. 2000, arXiv:astro-ph/0008187v1
- Cowan, N.B.; Ivezić, Z. 2008 ApJ674, L13
- Croft, R.A.C. & Efstathiou, G. 1994, MNRAS, 267, 390
- Croton D.J. et al., 2005, MNRAS, 356, 1155
- Croton D.J., Gao, L. & White, S.D.M 2007, MNRAS, 374, 1303
- Daley, D. J. & Vere-Jones, D. 2002 *An Introduction to the Theory of Point Processes, Volume I:Elementary Theory and Methods*, 2nd edition, Springer-Verlag: New York
- Davis, M., Huchra, J. Latham, D.W. & Tonry, J. 1982, ApJ, 253, 445
- de Berg, M., van Kreveld, M., Overmars, M., and Schwarzkopf, O. (1997), *Computational Geometry: Algorithms and Applications*, Springer-Verlag: New York

- Dekel, A. & West, M.J. 1985, ApJ, 288, 411
- DeSieno, D. 1988, “Adding a conscience to competitive learning”, IEEE International Conference on Neural Networks vol. 1, pp. 11171124
- Gregory, S.A. & Thompson, L.A. 1978, ApJ, 222, 784
- Holmberg, E. 1937 Ann. Obs. Lund No. 6, 1937
- de Vaucouleurs, G. 1953, AJ, 58, 30
- de Vaucouleurs, G. 1958, AJ, 63, 252
- de Vaucouleurs, G. 1975, in *Stars and Stellar Systems*, vol. 9, ed. A. Sandage, M. Sandage, and J. Kristian, (Chicago: University of Chicago Press)
- Daley, D. J. & Vere-Jones, D. 1999, *Formation of Structure in the Universe*, Cambridge University Press.
- Diehl, S. & Statler, T. 2006, MNRAS, 368, 497
- Doroshkevich, A. et al. 1996, MNRAS, 283, 1281
- Doroshkevich, A.G., Tucker, D.L., Fong, R., Turchaninov, V. & Lin, H. 2001, MNRAS, 322, 369
- Doroshkevich, A., Tucker, D.L., Allam, S. & Way, M.J. 2004, A&A, 418, 7
- Dressler, A. 1980, ApJ, 236, 351
- Einasto, J., Klypin, A.A., Saar, E., Shandarin, S.F. 1984, MNRAS, 206, 529
- Ebeling, H. & Wiedenmann, G., Phys. Rev. E, 47, 704
- Efstathiou, G. & Eastwood, J.W. 1981, MNRAS, 194, 503
- Elyiv, A., Melnyk, O. & Vavilova, I. 2009, MNRAS, 394, 1409-1418.
- Gamow, G. 1954, Proc. Natl. Acad. Sci. 40, 480
- Gazis, P.R. & Scargle, J.D. 2008, arXiv:0802.0861v1
- Gazis, P.R., Levit, C. & Way, M.J. 2010, arXiv:1008.2205v2
- Geller, M.J. & Huchra, J.P. 1983, ApJS, 52, 61

- Giovanelli, R. & Haynes, M.P. 1991, ARA&A, 29, 499
- Gomez, P.L., et al. 2003, ApJ, 584, 210
- Gott, J.R., Turner, E.L. & Aarseth, S.J. 1979, ApJ, 234, 13
- Gott, J.R., Melott, A.L. & Dickinson, M. 1986, ApJ, 306, 341
- Gott, J.R., Weinberg, D.H. & Melott, A.L 1987, ApJ, 319, 1
- Gott, J.R., Yun-Young, C., Park, C. & Kim, J 2009, ApJ, 695, L45
- Gray, A. & Moore, A. 2003, Rapid Evaluation of Multiple Density Models, in C. M. Bishop and B. J. Frey (eds), Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics, Jan 3-6, 2003, Key West, FL.
- Gray, A. & Moore, A. 2003, Nonparametric Density Estimation: Toward Computational Tractability, in Proceedings of the 2003 SIAM International Conference on Data Mining, May 1-3, 2003, San Francisco, CA.
- Groth, E.J. & Peebles, P.J.E. 1977 ApJ, 217, 385
- Hahn, O., Porciani, C., Carollo, C.M. & Dekel, A. 2007, MNRAS, 375, 489
- Hahn, O., Carollo, C.M., Porciani, C. & Dekel, A. 2007, MNRAS, 381, 41
- Hamilton, A.J.S., Gott, R.J. & Weinberg, D.H. 1986, ApJ309, 1
- Herschel, J.F.W. 1847, "Results of astronomical observations made during the years 1834, 5, 6, 7, 8, at the Cape of Good Hope; being the completion of a telescopic survey of the whole surface of the visible heavens, commenced in 1825," Phil. Trans. 137, 1
- Herschel, W. 1784, "Account of some observations tending to investigate the construction of the heavens, " Phil. Trans. 74, 437
- Hogg, D. W., Blanton, M. R., Eisenstein, D. J., Gunn, J. E., Schlegel, D. J., Zehavi, I., Bahcall, N. A., Brinkmann, J., Csabai, I., Schneider, D. P., Weinberg, D. H., York, D. G., 2003, ApJ, 585, L5
- Hubble, E.P. 1925, Pub. Am. Astr. Soc., 5, 261
- Hubble, E.P. 1934, ApJ, 79, 8
- Hubble, E.P. 1936, The Realm of the Nebulae (New Haven: Yale Univ. Press)

- Huchra, J., Davis, M., Latham, D., Tonry, J. 1983, ApJS, 52, 89
- Huchra, J. & Gellar, M.J. 1982, ApJ, 257, 423
- Huggins, W. 1864, “On the spectra of some nebulae,” Phil. Trans. 154, 437
- Hsu, A. & Halgamuge, S.K. (2003), “Enhancement of topology preservation and hierarchical dynamic self-organising maps for data visualisation.” *Int. J. Approximate Reasoning*, **23**, pp. 259-279.
- Icke, V. & van de Weygaert, R. 1987, A&A, 184, 16
- Ikeuchi, S. & Turner, E.L. 1991, MNRAS, 250, 519
- Ivezić, Z., Vivas, A.K., Lupton, R.H., Zinn, R. 2005, AJ, 129, 1096
- Ivezic, Z., Tyson, J.A., Allsman, R., Andrew, J., Angel, R., et al 2008, arXiv:0805.2366v1
- Jackson, B., Scargle, J.D., Barnes, D., Arabhi, S., Alt, A., Gioumoussis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., Tun Tao Tsai, An algorithm for optimal partitioning of data on an interval, IEEE Signal Processing Letters, 2005, Vol. 12, 105- 108.
- Jackson, B, Scargle, J., Cusanza, C., Barnes, D., Kanygin, D., Sarmiento, R., Subramaniam, S., & Chuang, T. 2010, “Optimal Partitions of Data in Higher Dimensions”, submitted to Statistical Analysis and Data Mining
- James, J.B., Lewis, G.F. & Colless, M. 2007, MNRAS, 375, 128
- Jones, B.J.T, van de Weygaert, R. & Aragón-Calvo, M.A. 2010, Submitted to MNRAS, arXiv:1001.4479
- Kaiser, N. et al. 2002, Proc. Of the SPIE, 4836, 154
- Kauffmann, G., White, S D.M., Heckman, T.M., Menard, B., Brinchmann, J., Charlot, S., Tremonti, C., & Brinkmann, J. 2004, MNRAS, 353, 713
- Kim, R.S.J., Strauss, M.A., Bahcall, N.A., Gunn, J.E., Lupton, R.H., Vogeley, M.S., Schlegel, D., for the SDSS collaboration, “Finding Clusters of Galaxies in the Sloan Digital Sky Survey using Voronoi Tessellation”, Clustering at High Redshift Les Rencontres Internationales de l’IGRAP, ASP Conference Series, Vol. 3108, 1999, Eds., A. Mazure, O. LeFevre, V. Lebrun
- Kohonen, T., Self-Organization and Associative Memory, Springer-Verlag, Berlin, 1984.

- Krzewina, L.G. & Saslaw, W.C. 1996, MNRAS, 278, 869
- Kutoyants, Yu. A. 1998, *Statistical Inference for Spatial Poisson Processes*, Lecture Notes in Statistics, Number 134, Springer: New York
- Landy, S.D. & Szalay, A.S. 1993, ApJ, 412, 64
- Layzer, D.N. 1956, AJ, 61, 383
- Lemson, G. & Kauffmann, G. 1999, MNRAS, 302, 111
- Limber, D.N. 1953, ApJ, 117, 134
- Limber, D.N. 1954, ApJ, 119, 655
- Limber, D.N. 1957, ApJ, 125, 9
- Martinez, V.J. & Saar, E., 2001, *Statistics of the Galaxy Distribution*, Chapman and Hall/CRC: Boca Raton; ISBN 1584880848
- Martinez, V.J. et al. 2005, ApJ, 634, 744
- Matsuda, T. & Shima, E. 1984, Prog. Theor. Phys., 71, 855
- Melnyk, O. V., Elyiv, A. A., NS Vavilova, I. B. (2006)M Kinematika i Fizika Nebesnykh Tel. , 22, 283-296 (2006) astro-ph:0712.1297
- Merényi, E. 1998, “Self-Organizing ANNs for Planetary Surface Composition Research”, Proc. European Symposium on Artificial Neural Networks, ESANN98, Bruges, Belgium, 22-24 April, 1998, pp 197-202.
- Messier, C. 1781, “Catalogue des nebuleuses et des amas d’eoiles” In Connoissance des Temps Pour l’Annee Bissexile 1784, Paris, p.263
- Miller, C.J. et al. 2005, AJ, 130, 968
- Moore, B. et al. 1992, MNRAS, 256, 477
- Mowbray, A.G. 1938, PASP, 50, 275
- Neyman, J. & Scott, E.L. 1952, ApJ, 116, 144
- Neyman, J. & Scott, E.L. 1959, ‘Large scale organization of the distribution of galaxies.’ in Encyclopedia of Physics, ed. S. Flugge (Berlin: Spring-Verlag) 53, 416.

- Neyman, J. 1962, “Alternative stochastic models of the spatial distribution of galaxies,” in Problems of Extra-Galactic Research, ed. G.C. McVittie, London, Macmillan, p. 294
- Neyrinck, M.C. 2008, MNRAS, 386, 2101
- Neyrinck, M.C., Gnedin, N.Y. & Hamilton, A.J.S. 2005, MNRAS, 356, 1222
- Oemler, A. 1974, ApJ, 194, 1
- Okabe, A., Boots, B., Sugihara, K., Chiu, S.N., & Kendall, D. G. 2000, “Spatial Tessellations: Concepts and Applications of Voronoi Diagrams”, 2nd edition, John Wiley & Sons, Ltd. New York
- Oort, J.H. 1983, ARA&A, 21, 3730
- Pandey, B. & Bharadwaj, S. 2005, MNRAS, 357, 1068
- Park, C. & Gott, J.R. 1991, ApJ, 378, 457
- Papoulis, Athanasios 1965, *Probability, Random Variables, and Stochastic Processes*, McGraw-Hill Book Co.: New York
- Paredes, S., Jones, B.J.T & Martinez, V.J. 1995, MNRAS, 276, 1116
- Pearson, R.C. & Coles, P. 1995, MNRAS, 272, 231
- Peebles, P.J.E. & Hauser, M.G. 1974, ApJS, 28, 19
- Peebles, P.J.E. 1980, *The Large-Scale Structure of the Universe*, Princeton, N.J.: Princeton University Press
- Pizarro, D., Campusano, L.E., Clowes, R.G., Virgili, P., Hitschfeld-Kahler, N. & Sochting, I.K. 2006, “Clustering of 3D Spatial Points Using Maximum Likelihood Estimator over Voronoi Tessellations: Study of the Galaxy Distribution in Redshift Space”, Proc. of the 3rd Intl. Sym. on Voronoi Diagrams in Science and Engineering (ISVD’06), IEEE Computer Society
- Platen, E., van de Weygaert, R. & Jones, B.J.T 2007, MNRAS380, 551
- Postman, M. & Geller, M.J. 1984, ApJ, 281, 95
- Preparata, F. P. & Shamos, M. I. (1985), *Computational Geometry: An Introduction*, Springer Verlag: New York.
- Press, W.H. & Davis, M. 1982, ApJ, 259, 449

- Ramella, M., Pisani, A. & Geller, M. 1997, *AJ*, 113, 483
- Ramella, M., Nonino, M., Boschin, W., & Fadda, D. 1999, in *ASP Conf. Ser. 176, Observational Cosmology: The Development of Galaxy Systems*, ed. G. Giuricin, M. Mezzetti, & P. Salucci (San Francisco: ASP), 108, <http://arxiv.org/abs/astro-ph/9810124>;
- Ramella M., Boschin W., Fadda D., Nonino M. 2001, *A&A*, 368, 776
- Rapetti, D, Allen, S.W., Mantz, A. & Ebeling, H. 2009, arXiv:0911.1787v2
- Reiz, A. 1941 *Ann. Obs. Lund* No. 9, 1941
- Ritter, H., T. Martinez, K. Schulten, *Neural Computation and Self-Organizing Maps*, Addison-Wesley, Reading, Mass., 1992.
- Rubin, V.C. 1954, *Proc. Natl. Acad. Sci.* 40, 541
- Sandage, A., & Tammann, G.A. 1975, *ApJ*, 197, 265
- Santiago, B.X. & Strauss, M.A. 1992, *ApJ*, 387, 9
- Saslaw, William C. 2000, *The Distribution of the Galaxies: Gravitational Clustering in Cosmology*, Cambridge University Press: Cambridge
- Scargle, J. 1998, *ApJ*, 504, p.405.
- Scargle, J. 2002 “Bayesian blocks in two or more dimensions: Image segmentation and cluster analysis,” in *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, American Institute of Physics Conference Proceedings, Volume 617, pp. 163-173.
- Scargle, J., Norris, J., & Jackson, B. 2008 “Studies in Astronomical Time Series Analysis. VI. Optimal Segmentation: Blocks, Triggers, and Histograms,” in preparation
- Schaap, W.E. & van de Weygaert, R. 2000, *A&A*, 363, L29
- Schaap, W.E. 2007, “The Delaunay Tessellation Field Estimator”, Ph.D. Thesis, Groningen University
- Schlegel, D.J. et al. 2009, arXiv:0904.0468v3
- Scott, David W. 1992, *Multivariate Density Estimation: Theory, Practice and Visualization*, John Wiley & Sons, Inc.: New York
- Shandarin, S.F. 1983, *Sov Astr. Letters*, 9, 104

- Shane, C.D. & Wirtanen, C.A. 1967, Publ. Lick. Obs. 22, Part 1
- Shapley, H. 1933, PNAS, 19, 389
- Shectman, S.A. et al. 1996, ApJ, 470, 172
- Sheth, J.V., Sahni, V., Shandarin, S.F. & Sathyaprakash, B.S. 2003, MNRAS, 343, 22
- Sheth, J.V. & Tormen, G. 2004, MNRAS, 350, 1385
- Silverman, B.W. 1986, *Density Estimation for Statistics and Data Analysis*, (Chapman & Hall; reprinted in 1998 by CRC Press: Boca Raton)
- Slezak, E., Bijaoui, A. & Mars, G. 1990, A&A, 227, 301
- Slezak, E., de Lapparent, V. & Bijaoui, A. 1993, ApJ, 409, 517
- Snyder, Donald L. & Miller, Michael I. 1991, *Random Point Processes in Time and Space*, 2nd edition, Springer-Verlag: New York
- Sousbie, T., Columbi, S. & Pichon C. 2009, MNRAS, 393, 457
- Sousbie, T. 2010, submitted to MNRAS, arXiv:1009.4105
- Sousbie, T., Pichon, C. & Kawahara, H. 2010, submitted to MNRAS, arXiv:1009.4104
- Stein, M.L. 1997 in Feigelson E.D., Babu G.J., eds., *Statistical Challenges in Modern Astronomy II*. Springer-Verlag, New York, p.166
- Stoyan, D., Kendall, W.S. & Mecke, J. 1995, *Stochastic Geometry and Its Applications*, ed. John Wiley & Sons, Chichester
- Springel, V., et al. 2005, Nature, 435, 629
- Strauss, M. A., et al. 2002 AJ, 124, 1810
- Stril, A., Cahn, R., & Linder E.V. 2010, MNRAS, 404, 239
- Szapudi, I & Szalay, A.S. 1998, ApJ, 494, 41
- Totsuji, H. & Kihara, T. 1969, PASJ, 21, 221
- Turner, E.L. & Gott, J.R. 1976, ApJS, 32, 409
- Turner, E.L., Aarseth, S.J., Gott, J.R., Blanchard, N.T. & Mathieu, R.D. 1979, ApJ, 228, 684

- Ueda, H. & Itoh, M. 1997, PASJ, 49, 131
- van de Weygaert R. 1994, A&A, 283, 361
- van de Weygaert R. 2003, “The Cosmic Foam: Stochastic Geometry and Spatial Clustering across the Universe,” Invited contribution in Proceedings of Statistical Challenges in Modern Astronomy III, eds. E.D. Feigelson & G.J. Babu, Springer-Verlag, pp. 175-196
- van de Weygaert R. & Schaap, W. 2009, “The Cosmic Web: Geometric Analysis”, in Data Analysis in Cosmology, Lecture Notes in Physics, vol. 665, Eds V.J. Martinez, E. Saar, E. Martinez-Gonzalez, and M.-J. Pons-Bordera. Berlin: Springer, 2009., p.291-413
- van de Weygaert R. & Aragón-Calvo, M. 2003, “Geometry and Morphology of the Cosmic Web: Analyzing Spatial Patterns in the Universe,” Invited review ISVD09 (International Symposium on Voronoi Diagrams and Engineering), Copenhagen, Denmark. IEEE CPS, E3781, ed. F. Anton.
- Villmann, T., Der, R., Herrmann, M., & Martinetz, T., 1997, “Topology Preservation in Self-Organizing Feature Maps: Exact Definition and Measurement.” *IEEE Transactions on Neural Networks*, 8, pp. 256-266.
- Vogele, M.S., et al. 1994, ApJ, 420, 525
- Wright, T. 1750, “An Original Theory or New Hypothesis of the Universe” New York: Elsevier 1971
- York, D.G. et al. 2000, AJ, 120, 1579
- Yoshioka, S. & Ikeuchi, S. 1989, ApJ, 341, 16
- Zehavi, I. et al. 2002, ApJ, 571, 172
- Zehavi, I. et al. 2010, arXiv:1005.2413
- Zel’dovich, Ya. B. 1970, A&A5, 84
- Zel’dovich, Ya. B., Einasto, J., & Shandarin, S.F. 1982, Nature, 300, 407
- Zhang, Y., Springel, V. & Yang, X. 2010, arXiv:1006.3768
- Zwicky, F. 1957, PASP, 69, 518
- Zwicky, F., Wield, P., Herzog, E., Karpowicz, M., & Kowal, C.T. 1961-68, Catalogue of Galaxies and Clusters of Galaxies, 6 volumes. Pasadena, California Institute of Technology

